DOCUMENT RESUME

ED 386 487 TM 024 056

AUTHOR Longford, Nicholas T.
TITLE Logistic Regression with Random Coefficients.
INSTITUTION Educational Testing Service, Princeton, NJ. Program Statistics Research Project.
REPORT NO ETS-RR-93-20; ETS-TR-93-30
PUB DATE Mar 93
NOTE 76p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS Algorithms; *Estimation (Mathematics); *Maximum Likelihood Statistics; *Scoring; Simulation
IDENTIFIERS Covariance Structure Models; Linear Models; *Logistic Regression; *Random Coefficient Models

ABSTRACT

An approximation to the likelihood for the generalized linear models with random coefficients is derived and is the basis for an approximate Fisher scoring algorithm. The method is illustrated on the logistic regression model for one-way classification, but it has an extension to the class of generalized linear models and to more complex data structures, such as nested two-way classification. Both full and restricted maximum likelihood versions of this algorithm are defined. The estimators of the regression parameters coincide with the generalized estimating equations of S. L. Zeger and K. Y. Liang (1986) but an essentially different class of estimators for the covariance structure parameters is obtained. A simulation study explores the properties of the proposed estimators. Five tables, 12 figures, and an appendix of statistical analysis are included. (Contains 43 references.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 386 487

RR-93-20

Logistic Regression With Random Coefficients

Nicholas T. Longford
Educational Testing Service

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

PROGRAM STATISTICS RESEARCH

TECHNICAL REPORT NO. 93-30

Educational Testing Service
Princeton, New Jersey 08541

BEST COPY AVAILABLE

TM024056

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the members of the Research Statistics Group as well as their external and visiting statistical consultants.

Reproduction of any portion of a Program Statistics Research Technical Report requires the written consent of the author(s).

Logistic Regression With Random Coefficients

Nicholas T. Longford
Educational Testing Service

Program Statistics Research
Technical Report No. 93-30

Research Report No. 93-20

Educational Testing Service
Princeton, New Jersey 08541

March 1993

Copyright © 1993 by Educational Testing Service. All rights reserved.

LOGISTIC REGRESSION WITH RANDOM COEFFICIENTS

N. T. Longford
Educational Testing Service, Princeton, NJ

Abstract

An approximation to the likelihood for the generalized linear models with random coefficients is derived and is the basis for an approximate Fisher scoring algorithm. The method is illustrated on the logistic regression model for one-way classification, but it has an extension to the class of generalized linear models and to more complex data structures, such as nested two-way classification. Both full and restricted maximum likelihood versions of this algorithm are defined. The estimators of the regression parameters coincide with the generalized estimating equations of Zeger and Liang (1986) but an essentially different class of estimators for the covariance structure parameters is obtained. A simulation study explores the properties of the proposed estimators.

Some key words: covariance structure, Fisher scoring method, logistic regression, maximum likelihood, random coefficients.

Acknowledgements

I wish to acknowledge Ms. Kahn's review of Section 8.2. Neal Thomas introduced me to the problem discussed in Section 8.2. Gerald Shure and Eric Dey introduced me to the problem and the dataset analysed in Section 8.1. Neal Thomas, Steffen Lauritzen, and Murray Aitkin provided valuable and insightful comments on earlier versions of this paper. Neal Thomas and John Donoghue reviewed the manuscript submitted for the ETS Technical Report Series. I take responsibility for any errors that remain in the paper.

1 Introduction

Clustered observations arise in a wide variety of applications, including agricultural and animal breeding studies, econometrics, educational and medical research, and survey analysis. For regression of such data random coefficient models are usually considered, so as to take account of, or to make inference about, the between-cluster variation.

There are several well-established and well-researched computational algorithms for fitting random coefficient models with the usual normal assumptions; for a comprehensive review of earlier developments, see Harville (1977). In this paper we concentrate on a logistic regression model for correlated binary outcomes, although the methods discussed have direct extensions for binomial data with link functions other than logistic, and more generally, to the entire class of generalized linear models.

Our development is similar to the generalized estimating equations (GEE) of Liang and Zeger (1986) and Zeger and Liang (1986), although we establish a more direct connection between the generalized linear models with random coefficients and the corresponding computational algorithms. For example, our computational algorithm is based on an approximation to the log-likelihood, and estimation procedures have their full and restricted (approximate) maximum likelihood versions. We propose a new estimator for the covariance structure parameters which is a generalization of the maximum likelihood estimator in the normal case.

Section 2 describes random coefficient models, and gives a brief summary of computational algorithms for fitting such models for binary data. Section 3 discusses a maximum likelihood procedure using Gaussian quadrature, and Section 4 presents a procedure based on an approximation to the log-likelihood. Extensions of this method for more complex data structures, and for other generalized linear models are indicated in Section 5. Formal derivation of the approximation to the log-likelihood for correlated observations with a distribution from the exponential family is given in the Appendix. Section 6 describes an adaptation of the exact and approximate procedures for restricted maximum likelihood. In Section 7 we study the bias of the generalized least squares estimator for the regression parameters. We establish a relationship between the bias of the ordinary least squares estimator in the normal model with random coefficients and our approximation to the bias in the logistic regression.

The methods are illustrated and compared on two examples (Section 8), and the properties of the associated estimators are explored in a simulation study (Section 9). We conclude that the generalized least squares method provides a satisfactory estimator except when there is extremely large between-cluster variation.

We focus on logistic regression and binary data, since they are frequently considered in practice, and binary data represent an extreme form of departure from normality. Random coefficient methods for clustered observations

provide a parsimonious summary for between-cluster differences.

2 Logistic regression models

We consider the logistic regression model

$$\text{logit}\{P(y_{ij} = 1 \mid \delta_j)\} \;=\; \mathbf{x}_{ij}\boldsymbol{\beta} + \sigma\delta_j \tag{1}$$

for binary outcomes $\{y_{ij}\}$ with a one-way layout structure, that is, $i = 1,\ldots,n_j$ and $j = 1,\ldots N_2$, and explanatory variables $\mathbf{x}$. The number of elementary-level units is $N = n_1 + \ldots + n_{N_2}$. The regression parameters $\boldsymbol{\beta}$ and the variance parameter σ^2 may be known or unknown, and $\sigma^2 \geq 0$. Each regressor x may be defined either for individuals (elementary observations) or for the clusters; in the latter case the index i is redundant, since such a variable is constant within clusters. We assume that $\{\delta_j\}$ are a random sample from the standard normal distribution, $\delta_j \sim \text{N}(0,1)$, i.i.d. The model (1) has a straightforward extension to models with more complex patterns of between-cluster variation:

$$\text{logit}\{P(y_{ij} = 1 \mid \boldsymbol{\delta}_j)\} \;=\; \mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\delta}_j\,, \tag{2}$$

where $\boldsymbol{\Sigma}$ is a non-negative definite variance matrix (the possible non-uniqueness of the square root in (2) is immaterial), and $\boldsymbol{\delta}_j \sim \text{N}_r(\mathbf{0},\mathbf{I})$. The variables

$\mathbf{z}$ are usually a subset of the variables $\mathbf{x}$, and both contain the intercept $\mathbf{1} = (1, \ldots, 1)^{\mathsf{T}}$. This conforms with similar conventions in analysis of covariance.

Extensions of the model (2) for data with a multi-way layout are straightforward. For example, for the nested two-way layout (clusters within areas) it is natural to consider the model

$$\text{logit}\{P(y_{ijk} = 1 \mid \boldsymbol{\delta}_{jk,1}, \boldsymbol{\delta}_{k,2})\} \;=\; \mathbf{x}_{ijk}\boldsymbol{\beta} + \mathbf{z}_{ijk,1}\boldsymbol{\Sigma}_1^{\frac{1}{2}}\boldsymbol{\delta}_{jk,1} + \mathbf{z}_{ijk,2}\boldsymbol{\Sigma}_2^{\frac{1}{2}}\boldsymbol{\delta}_{k,2} \quad (3)$$

where $i = 1, \ldots, n_{jk}^{(1)}, j = 1, \ldots, n_k^{(2)}$ and $k = 1, \ldots, N_3$ are the indices for the elementary observation within a cluster, for the cluster within an area, and for the area, respectively, and the components of the random vectors $\boldsymbol{\delta}_{jk,1}$ and $\boldsymbol{\delta}_{k,2}$ are (univariate) independent N(0,1) random variables.

Further generalization of the models (1) – (3) involves spanning them over the class of generalized linear models (Nelder and Wedderburn, 1972, and McCullagh and Nelder, 1990): Conditionally on the random vectors $\{\boldsymbol{\delta}_j\}$ the outcomes $\{y_{ij}\}$ have a specified distribution (e.g., Poisson, gamma, or t), and

$$h\{\mathbf{E}(y_{ij} \mid \boldsymbol{\delta}_j)\} \;=\; \mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\delta}_j, \quad (4)$$

where h is a link function. This model is a generalization of (2); the generalization of (3) is analogous. In addition to the design matrix for variation,

given by the variables $\mathbf{z}$ and the matrix $\boldsymbol{\Sigma}$, it suffices to specify only the link function h and the dependence of the conditional variance $\mathrm{var}(y_{ij}|\boldsymbol{\delta}_j)$ on the conditional mean $\mathbf{E}(y_{ij}|\boldsymbol{\delta}_j)$.

The model (4) with the identity link $h(x) = x$ and the normal distribution,

$$y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\delta}_j + \epsilon_{ij}, \tag{5}$$

where $\epsilon_{ij} \sim \mathrm{N}(0, \sigma_0^2)$, represents a special case. We will refer to (5), with $\sigma_0^2 = 1$, as the (normal) *parent* model for (2), and similarly define the parent models for (3) and (4).

There are several methods for maximum likelihood estimation in the model (5). For earlier results see Hartley and Rao (1967), Patterson and Thompson (1971), and for comprehensive reviews, Searle (1971) and Harville (1977). Dempster, Laird and Tsutakawa (1981) describe an application of the EM algorithm, Jennrich and Schluchter (1986) and Longford (1987) give details of Newton-Raphson and Fisher scoring procedures, and Goldstein (1986) describes a generalized least squares algorithm. The EM algorithm tends to require a substantially higher number of iterations than the other methods, even when it is combined with routines designed to accelerate convergence; see Thompson and Meyer (1986) and Lindstrom and Bates (1988) for more detailed discussion. An iteration of the EM algorithm and one of

the Fisher scoring algorithm are of comparable complexity, since both require computation of certain quadratic forms in the inverses of the unconditional variance matrices $\mathrm{var}(\mathbf{y}_j) = \sigma_0^2\mathbf{I} + \mathbf{Z}_j\boldsymbol{\Sigma}\mathbf{Z}_j^T$, where $\mathbf{y}_j = (y_{1j}, \ldots, y_{n_j,j})^\top$ and $\mathbf{Z}_j = (\mathbf{z}_{1j}^\top, \ldots, \mathbf{z}_{n_j,j}^\top)^\top$. Lange and Laird (1989) describe some special cases for which maximum likelihood solutions can be expressed in a closed form. An important subset of these cases can be described as having balanced design for the random part, i.e., $\mathbf{Z}_j$ are identical across the clusters.

Relative simplicity of the algorithms for fitting the normal model (5) can be attributed to the existence of closed form expressions for the conditional moments of the random terms given the outcomes $\{y_{ij}\}$ (for the EM algorithm), and of the log-likelihood and its partial derivatives (for Fisher scoring, Newton-Raphson or generalized least squares methods).

For the logistic regression model (2) the joint likelihood for $\{y_{ij}\}$ involves normal integrals;

$$l = \sum_j \log \int_{\mathbf{R}^r} \cdots \int P_j(\boldsymbol{\delta}_j)\boldsymbol{\Phi}_r(\boldsymbol{\delta}_j)\mathrm{d}\boldsymbol{\delta}_j \quad (= \sum_j l_j), \tag{6}$$

where $P_j(\boldsymbol{\delta}_j) = \prod_{i=1}^{n_j}\{p_{ij}(\boldsymbol{\delta}_j)\}^{y_{ij}}\{1-p_{ij}(\boldsymbol{\delta}_j)\}^{1-y_{ij}}$ is the conditional likelihood for cluster j given $\boldsymbol{\delta}_j$, $p_{ij}(\boldsymbol{\delta}_j) = \mathrm{logit}^{-1}(\mathbf{x}_{ij}\boldsymbol{\beta}+\mathbf{z}_{ij}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\delta}_j)$, and $\boldsymbol{\Phi}_r$ is the density function of $\mathbf{N}_r(\mathbf{0}, \mathbf{I})$. Direct maximization of (6) can be accomplished by using

Gaussian quadrature, although with three or more dimensional integrals this may be computationally very extensive.

The EM algorithm avoids evaluation of the integrals in (6), although the conditional moments of the random effects $\boldsymbol{\delta}_j$ involve integrals, and so there appears to be no gain in computational efficiency; on the contrary, owing to slow convergence of the EM algorithm the latter integrals have to be evaluated a larger number of times than the integrals in a direct maximization routine. Laird and Ware (1982) and Stiratelli, Laird and Ware (1984) replace the conditional expectations required for the EM algorithm by the conditional modes, thus reducing the computational load somewhat. See also Anderson and Aitkin (1985) for discussion of the EM algorithm in this context.

The intractable form of the log-likelihood (6) has until recently effectively discouraged application of the normal-mixture model (2), and the beta-binomial model (Williams, 1982), in which variation of the within-cluster (i.e., conditional) probabilities is modelled by a beta distribution, has been preferred. Simplification of the corresponding likelihood takes place since the beta distribution is the conjugate for the binomial distribution. A distinct disadvantage of this approach for applications where between-cluster variation is of substantive interest is that the scale of the mixing beta distribution is difficult to relate to the more familiar logit or probit scales. Also, the method is specific to binomial data, although other familiar distributions have their conjugates. Rosner (1984) and Prentice (1986) discuss and extend

the results of Williams (1982). Breslow (1984) adapts the idea for clustered Poisson data.

Bonney (1987) and Connolly and Liang (1988) define a class of logistic regression models for non-independent observations by the conditional distributions of the individual outcomes given the rest of the outcomes in the cluster (or a subset thereof). Although this development is most natural for times series or longitudinal data it is equally suitable for situations with a symmetric pattern of dependence.

The generalized estimating equations approach (GEE) of Zeger and Liang (1986) and Liang and Zeger (1986) is based on a generalized least squares type estimator for the regression parameters

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{y}, \tag{7}$$

where $\mathbf{X}$ is the design matrix, consisting of the rows $\mathbf{x}_{ij}$, in lexicographic order, $\mathbf{y}$ is the corresponding vector of outcomes, and $\mathbf{V}$ is a variance matrix, a function of certain parameters. Zeger and Liang (1986) propose naive estimators for these parameters; Prentice (1988) discusses a class of more efficient estimators. Thus model fitting by the GEE approach involves iterations of (7), with the parameters in $\mathbf{V}$ replaced by their current estimates, and an updating of the parameters involved in $\mathbf{V}$.

In contrast to the other methods reviewed above, GEE approach does not

arise from a model that describes how the data are generated. This renders assessment of the assumed model (model checking) difficult, especially for small samples. Nevertheless the GEE approach has two important virtues: computational simplicity, and that it caters for a general class of distributional assumptions as well as link functions. A similar framework, based on quasilikelihood, is proposed by McCullagh and Nelder (1990).

For logistic regression Zeger and Liang (1986) propose a general covariance structure

$$\mathbf{V} = \{\mathbf{R}(\boldsymbol{\rho})\}^{\frac{1}{2}}\mathbf{V}_0\{\mathbf{R}(\boldsymbol{\rho})\}^{\frac{1}{2}}, \tag{8}$$

where $\mathbf{V}_0$ is the diagonal matrix, $\mathbf{V}_0 = \text{diag}\{p_{ij}(1-p_{ij})\}$, $p_{ij} = \text{logit}^{-1}(\mathbf{x}_{ij}\boldsymbol{\beta})$, and $\mathbf{R}(\boldsymbol{\rho})$ is a block-diagonal correlation matrix, with blocks $\mathbf{R}_j(\boldsymbol{\rho})$ corresponding to the clusters. The simplest non-trivial choice for $\mathbf{R}(\boldsymbol{\rho})$ is the equicorrelation structure,

$$\mathbf{R}_j(\rho) = (1-\rho)\mathbf{I}_{n_j} + \rho\mathbf{J}_{n_j}, \tag{9}$$

where $\mathbf{I}_n$ and $\mathbf{J}_n$ are the identity matrix and the matrix of ones, of size $n \times n$, respectively. Zeger and Liang (1986) and Zhao and Prentice (1990) discuss estimation procedures for ρ. Note that $\mathbf{R}(\rho)$ in (9) is non-negative definite for $\rho > -1/\max(n_j - 1)$; negative correlations can be realized when there is

an upper limit for the cluster sizes.

Alternative parametrizations for $\mathbf{R}$ include sequential dependence, e.g., tridiagonal $\mathbf{R}_j$ or tridiagonal $\mathbf{R}_j^{-1}$, applicable for longitudinal analysis; see Zeger, Liang, and Albert (1988), Zeger and Qadish (1988) and Zhao and Prentice (1990) for examples.

An exact maximum likelihood procedure for the model (1) has been proposed by Anderson and Aitkin (1985), and in a more general context (two-way nested layout) by Im and Gianola (1988). In Section 3 we describe an alternative to these methods based on the Newton-Raphson method. All these algorithms rely on numerical integration.

3 Exact maximum likelihood

In principle, evaluation of the integrals in (6), as well as of their partial derivatives, can be accomplished by Gaussian quadrature. This is feasible in practice for models with simple structure of between-cluster variation, that is, with a variance matrix $\boldsymbol{\Sigma}$ in (2) of low dimension, and then direct maximization of (6) is relatively straightforward. For the model (1) we have

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_j \exp(-l_j) \int_{-\infty}^{+\infty} P_j(\delta)\mathbf{s}_j(\delta)\Phi(\delta)d\delta, \qquad (10)$$

and

$$\frac{\partial l}{\partial \sigma} = \sum_j \exp(-l_j) \int_{-\infty}^{+\infty} P_j(\delta) s_{j1}(\delta) \delta \Phi(\delta) d\delta, \tag{11}$$

where l_j is the j^{th} summand defined in (6),

$$\mathbf{s}_j(\delta) = \sum_i \{y_{ij} - p_{ij}(\delta)\} \mathbf{x}_{ij}$$

and s_{j1} is the first component of $\mathbf{s}_j$ (corresponding to the intercept **1**). The second-order partial derivatives of (6) are

$$-\frac{\partial^2 l}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}^\top} = \sum_j \frac{\partial l_j}{\partial \boldsymbol{\beta}} \frac{\partial l_j}{\partial \boldsymbol{\beta}^\top} + \sum_j \exp(-l_j) \int_{-\infty}^{+\infty} P_j(\delta) \left\{ \mathbf{X}_j^\top \mathbf{W}_j(\delta) \mathbf{X}_j - \mathbf{s}_j(\delta) \mathbf{s}_j^\top(\delta) \right\} \Phi(\delta) d\delta, \tag{12}$$

$$-\frac{\partial^2 l}{\partial \sigma\, \partial \boldsymbol{\beta}^\top} = \sum_j \frac{\partial l_j}{\partial \sigma} \frac{\partial l_j}{\partial \boldsymbol{\beta}^\top} + \sum_j \exp(-l_j) \int_{-\infty}^{+\infty} P_j(\delta) \left\{ \mathbf{X}_j^\top \mathbf{W}_j(\delta) \mathbf{z}_j - \mathbf{s}_j(\delta) s_{1j}^\top(\delta) \right\} \delta\, \Phi(\delta) d\delta, \tag{13}$$

and

$$-\frac{\partial^2 l}{(\partial \sigma)^2} = \sum_j \left(\frac{\partial l_j}{\partial \sigma} \right)^2$$

$$+ \sum_j \exp(-l_j) \int_{-\infty}^{+\infty} P_j(\delta) \left\{ \mathbf{z}_j^T \mathbf{W}_j(\delta) \mathbf{z}_j - s_{1j}^2(\delta) \right\} \delta^2 \, \Phi(\delta) d\delta, \tag{14}$$

where $\mathbf{W}_j(\delta)$ is the diagonal matrix of the conditional variances, given $\delta_j = \delta$, $w_{ij}(\delta) = p_{ij}(\delta)\{1 - p_{ij}(\delta)\}, i = 1, \ldots, n_j$, $\mathbf{X}_j$ is formed by vertical stacking of the row vectors $\mathbf{x}_{ij}$, and $\mathbf{z}_j$ is the $n_j \times 1$ vector of ones. The expressions (6) and (10) – (14) can be approximated by Gaussian quadrature and used in a Newton-Raphson maximization procedure. The generalized least squares solution, which corresponds to $\sigma = 0$, is a suitable starting solution. An arbitrary positive number can be used as the starting solution for σ (or σ^2). In the examples discussed in Section 8 we set initially $\hat{\sigma} = 1$, usually much larger than the maximum likelihood solution; a more judicious choice for the starting $\hat{\sigma}$ would save not more than one Newton-Raphson iteration. For several model fits for these datasets, and for the number of quadrature points in the range 3 – 11, the maximization procedure required 3 – 5 GLS iterations and 3 – 6 further iterations based on (10) – (14).

It appears that 5 quadrature points are sufficient for data with moderate number of clusters (say, up to 100), but for larger datasets a higher number of quadrature points is required, although 9 points are sufficient even for data with 2,000 clusters. Empirical evidence for these observations is provided by the analysis of two datasets (Section 8) as well as by the simulation study reported in Section 9.

4 An approximate Fisher scoring algorithm

We denote the linear predictor $\theta_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta}$ and for brevity write $p_{ij} = p_{ij}(0)$, $w_{ij} = p_{ij}(1-p_{ij})$, $w_{+j} = \sum_i w_{ij}$, $e_{ij} = (y_{ij} - p_{ij})/w_{ij}$, $\mathbf{e}_j = (e_{1j}, \ldots, e_{n_j j})^\mathsf{T}$, $\mathbf{w}_j = (w_{1j}, \ldots, w_{n_j j})^\mathsf{T}$. Note that e_{ij} is the generalized residuals familiar from the generalized least squares method.

The conditional log-likelihood for the cluster j has the Taylor expansion around $\delta = 0$

$$\begin{aligned} \log\{P_j(\delta)\} &\approx \log\{P_j(0)\} + \sigma\delta\mathbf{e}_j^\mathsf{T}\mathbf{w}_j - \frac{1}{2}(\sigma\delta)^2 w_{+j} \\ &\quad - \ldots - \frac{(\sigma\delta)^k}{k!}\sum_i \frac{\partial^{k-1} p_{ij}(\delta)}{(\partial\delta)^{k-1}}|_{\delta=0}\,. \end{aligned} \tag{15}$$

If all but the first three terms of this expansion are ignored, we obtain

$$\begin{aligned} l &\approx \sum_j \log\{P_j(0)\} - \frac{N_2}{2}\log(2\pi\sigma^2) \\ &\quad + \sum_j \log\int_{-\infty}^{+\infty} \exp\left(\sigma\delta\mathbf{e}_j^\mathsf{T}\mathbf{w}_j - \frac{\delta^2 g_j}{2}\right) d\delta \\ &= \sum_j \log\{P_j(0)\} - \frac{1}{2}\sum_j \log\, g_j + \frac{\sigma^2}{2}\sum_j \frac{(\mathbf{e}_j^\mathsf{T}\mathbf{w}_j)^2}{g_j}\,, \end{aligned} \tag{16}$$

where $g_j = 1 + \sigma^2 w_{+j}$.

An approximate maximum likelihood estimator for the parameters $\boldsymbol{\beta}$

and σ^2 can be defined as the maximizer of (16) in the parameter space $(-\infty, +\infty)^p \times [0, +\infty)$. If we ignore the dependence of $\mathbf{w}_j$ on $\boldsymbol{\beta}$ we have

$$\frac{\partial l}{\partial \boldsymbol{\beta}} \approx \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{e}, \tag{17}$$

and

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \approx \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X}, \tag{18}$$

where $\mathbf{X}$ and $\mathbf{e}$ are formed by stacking of $\{\mathbf{X}_j\}$ and $\{\mathbf{e}_j\}$, respectively, and $\mathbf{V}$ is the block-diagonal matrix with blocks $\mathbf{V}_j = \mathbf{W}_j^{-1} + \sigma^2 \mathbf{J}_{n_j}$, and $\mathbf{W}_j = \mathbf{W}_j(0)$. See Longford (1988) for derivation of the approximations (17) and (18). The approximations imply the generalized least squares estimator (7), although with a parametric form for $\mathbf{V}$ different from (9): Whereas in the GEE approach, (9), equal *correlations* of the Pearson residuals are assumed, our development leads to equal *covariances* of the generalized residuals. We will refer to the estimator based on the latter as the AML (approximate maximum likelihood) estimator. Note that some improvement in the formulae (17) and (18) can be achieved by exact differentiation of (16).

If the linear predictor $\theta = \mathbf{x}\boldsymbol{\beta}$ is constant then the GEE and AML estimators of $\boldsymbol{\beta}$ $(= \beta_1)$ coincide because the corresponding variance matrices $\mathbf{V}$ are identical. In this case the parameters ρ and σ^2 are related by the identity $\rho = \sigma^2/(w + \sigma^2)$, where w is the common value of all w_{ij}. In general, the two parametrizations are more different the larger the variation of the

linear predictors θ_{ij}. For independent observations ($\rho = 0$, or $\sigma = 0$) both approaches result in the generalized least squares (GLS) estimator. Thus we can expect the GEE and AML estimators to perform well for small departures from independence of the observations. However, for high correlations (covariances) the variance matrices $\mathbf{V}$ used for obtaining these two estimators have unrealistic properties. For GEE, contrary to expectations, the modelled variances of the observations (the diagonal of $\mathbf{V}$) do not depend on ρ; the within-cluster correlation does not inflate the variances of the observations. For AML these variances are inflated by $(\sigma w_{ij})^2$, and for large σ some of the variances $w_{ij}(1 + \sigma^2 w_{ij})$ may be greater than $\frac{1}{4}$, the maximum variance of binary data. Thus caution should be exercised when interpreting fitted variances for the observations based on the estimated correlation $\hat{\rho}$ or the variance $\hat{\sigma}^2$.

The first-order partial derivative of (16) with respect to σ^2 is

$$\frac{\partial l}{\partial \sigma^2} \approx -\frac{1}{2}\sum_j \frac{w_{+j}}{g_j} + \frac{1}{2}\sum_j \left(\frac{\mathbf{e}_j^\top \mathbf{w}_j}{g_j}\right)^2, \qquad (19)$$

and the second-order partial derivative is

$$\frac{\partial^2 l}{(\partial \sigma^2)^2} \approx \frac{1}{2}\sum_j \frac{w_{+j}^2}{g_j^2} - \sum_j (\mathbf{e}_j^\top \mathbf{w}_j)^2 \frac{w_{+j}}{g_j^3}. \qquad (20)$$

Using the approximate identity $\mathrm{var}(\mathbf{e}_j^\top \mathbf{w}_j) = w_{+j} g_j$ we obtain

$$-\mathbf{E}\left\{\frac{\partial^2 l}{(\partial\sigma^2)^2}\right\} \approx \frac{1}{2}\sum_j \left(\frac{w_{+j}}{g_j}\right)^2, \tag{21}$$

and by similar operations it can be shown that $\mathbf{E}\{\partial^2 l/(\partial\boldsymbol{\beta}\partial\sigma^2)\} \approx \mathbf{0}$.

5 Extensions

In this Section we consider extensions for the AML approach parallel to the extensions of the basic GEE model (8) - (9).

First, the development (15) - (16) for the general model (2) yields the approximation

$$\begin{aligned} l \;&\approx\; [-N\,\log(2\pi) - \log\,\det(\mathbf{G}_j) + 2\sum_j \log\{P_j(0)\} \\ &+\; \sum_j \mathbf{e}_j^\top \mathbf{W}_j \mathbf{Z}_j\, \boldsymbol{\Sigma}\mathbf{G}_j^{-1}\, \mathbf{Z}_j^\top \mathbf{W}_j \mathbf{e}_j]/2, \end{aligned} \tag{22}$$

where $N = n_1 + \ldots + n_{N_2}$ is the sample size, $\mathbf{G}_j = \mathbf{I} + \mathbf{Z}_j^\top \mathbf{W}_j \mathbf{Z}_j \boldsymbol{\Sigma}$, and so it corresponds to the choice $\mathbf{V}_j = \mathbf{W}_j^{-1} + \mathbf{Z}_j \boldsymbol{\Sigma} \mathbf{Z}_j^\top$ in (17) and (18). Times series patterns of dependence can be implemented only by specifying the form of $\mathbf{V}_j$, in complete analogy with the GEE approach.

5.1 Nested two-way layout

Suppose the individual observations, indexed ijk, are in clusters jk, and the clusters are contained in *areas* $k = 1, 2, \ldots, N_3$. We will use the notation of

the previous sections, with the additional subscript k denoting the area. For the model (3) let

$$\mathbf{V}_{jk,1} = \mathbf{W}_{jk}^{-1} + \mathbf{Z}_{jk,1}\boldsymbol{\Sigma}_1\mathbf{Z}_{jk,1}^{\top}$$

and

$$\mathbf{V}_{k,2} = \text{diag}(\mathbf{V}_{jk,1}) + \mathbf{Z}_{k,2}\boldsymbol{\Sigma}_2\mathbf{Z}_{k,2}^{\top}. \tag{23}$$

Note that $\mathbf{V} = \text{diag}(\mathbf{V}_{k,2})$ would be the natural choice for the variance matrix in (17) and (18). Appendix contains a derivation of the following approximation for the density for the nested two-way layout in the context of the generalized linear models,

$$\begin{aligned} l \approx\ & \frac{1}{2}\{ - N\ \log(2\pi) - \sum_k \sum_j \log\ \det\ \mathbf{G}_{jk,1} - \sum_k \log\ \det\ \mathbf{G}_{k,2} \\ +\ & \sum_k \mathbf{e}_k^{\top}\mathbf{V}_{k,2}^{-1}\mathbf{Z}_{k,2}\ \boldsymbol{\Sigma}_2\mathbf{G}_{k,2}^{-1}\ \mathbf{Z}_{k,2}^{\top}\mathbf{V}_{k,2}^{-1}\mathbf{e}_k\} + \sum_k \sum_j \log\{P_j(0)\}, \end{aligned} \tag{24}$$

where $\mathbf{G}_{jk,1} = \mathbf{I} + \mathbf{Z}_{jk,1}^{\top}\mathbf{W}_{jk,1}\mathbf{Z}_{jk,1}\boldsymbol{\Sigma}_1$ and $\mathbf{G}_{k,2} = \mathbf{I} + \mathbf{Z}_{k,2}^{\top}\mathbf{V}_{k,2}^{-1}\mathbf{Z}_{k,2}\boldsymbol{\Sigma}_2$, and $\mathbf{e}_k$ is the vector of generalized residuals for the observations in the area k. Note that the inverse of the matrix $\mathbf{V}_{jk,1}$ can be expressed in terms of the matrix $\mathbf{G}_{jk,1}$.

Differentiation of (24) with respect to $\boldsymbol{\beta}$, while ignoring the dependence of $\mathbf{V}_{k,2}$, $\mathbf{G}_{jk,1}$ and $\mathbf{G}_{k,2}$ on $\boldsymbol{\beta}$, leads to the estimator given by (17) and (18),

with the blocks of $\mathbf{V}$ given by (23). For a general parameter ξ involved in one of the variance matrices, $\boldsymbol{\Sigma}_1$ or $\boldsymbol{\Sigma}_2$, we have

$$\frac{\partial l}{\partial \xi} \approx -\frac{1}{2}\mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \xi}\right) + \frac{1}{2}\sum_k \mathbf{e}_k^\top \mathbf{V}_{k,2}^{-1}\frac{\partial \mathbf{V}_{k,2}}{\partial \xi}\mathbf{V}_{k,2}^{-1}\mathbf{e}_k^\top \tag{25}$$

and

$$-\mathbf{E}\left(\frac{\partial^2}{\partial \xi_1 \partial \xi_2}\right) \approx \frac{1}{2}\mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \xi_1}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \xi_2}\right). \tag{26}$$

Proof of (25) and (26) is contained in the Appendix.

5.2 Generalized linear models

The Taylor expansion (15) can be applied in the much more general context of (conditional) generalized linear models defined by (4), or its extension for two-way layout. Details are given in the Appendix.

The normal model (5) with the identity link is a special case of these conditional GLM models. An exact maximum likelihood solution for this model is obtained by the iterative procedure defined by the equations (7), (17) – (19) and (21) with $\mathbf{V}_j = \sigma^2\mathbf{I} + \mathbf{Z}_j\boldsymbol{\Sigma}\mathbf{Z}_j^\top$. An important implication of this is that a maximum likelihood algorithm for fitting the normal model (5) can be adapted for the conditional GLM model (4) by replacing all the crossproducts required for evaluation of the normal log-likelihood (and of its partial derivatives) by its weighted versions, with the weights defined by the

applied link and (conditional) variance functions.

In the normal case the additional scale parameter $\sigma^2 = \mathrm{var}(\epsilon)$ can be estimated as the stationary point of the equation

$$N = \mathbf{e}^\top \mathbf{V}^{-1} \mathbf{e}. \tag{27}$$

It is advantageous to use the reparametrization $\boldsymbol{\Omega} = \sigma^{-2}\boldsymbol{\Sigma}$, in which case (27) is obtained by solving the normal equation for σ^2. Then $\hat{\sigma}^2 = \mathbf{e}^\top \hat{\mathbf{U}}^{-1}\mathbf{e}$, where $\mathbf{U} = \mathbf{I} + \mathrm{diag}(\mathbf{Z}_j \boldsymbol{\Sigma} \mathbf{Z}_j^\top)$ does not depend on σ^2.

6 Restricted maximum likelihood

In the normal case the maximum likelihood estimator for the variance matrix $\boldsymbol{\Sigma}$ is known to be biased, and an unbiased estimator is obtained by maximizing the likelihood corresponding to the $N - p$ error contrasts orthogonal to the regressor space; see Patterson and Thompson (1971), or Harville (1977), for detailed discussion. Harville (1974) has derived an explicit form for this *restricted* log-likelihood (RML). The *full* and restricted log-likelihoods differ by a constant and the term

$$R = -\frac{1}{2} \log \left\{ \det \left(\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X} \right) \right\}. \tag{28}$$

The formulae for the Fisher scoring algorithm can be straightforwardly ad-

justed for RML.

Stiratelli, Laird and Ware (1984) define the restricted log-likelihood for logistic mixed regression with random coefficients by integrating out the regression parameters β using a flat prior. RML is most important for models with relatively many regression parameters, and maximization of the likelihood may be computationally demanding exactly in such cases.

We propose to approximate the restricted log-likelihood by the adjustment (28), with the variance matrix $\mathbf{V}$, given by (17) or (23), as appropriate, to the respective log-likelihood (16) or (24), or indeed to the exact log-likelihood. This proposal has no theoretical foundation, it is based only on analogy with the normal case. The simulations reported in Section 9 show that the maximum likelihood estimator for the between-cluster variance in logistic regression is downward biased, and that the restricted maximum likelihood adjustment reduces its bias (in the simulated situation).

7 Bias of the GLS estimator

An important practical issue is related to the performance of the GLS estimator of the regression parameters in presence of positive between-cluster variation. In the context of the normal models this issue has been discussed by Holt and Scott (1982). The principal question is that of the bias of the estimators for the standard error. Since the approximation for the asymptotic information matrix for β in logistic regression has a form similar to its

exact counterpart for the normal regression, a discussion similar to that of Holt and Scott (1982) can be conducted. For illustration we assume a simple logistic regression model

$$\text{logit}\{P(y_{ij}=1|\delta_j)\} \;=\; \alpha+\beta x_{ij}+\sigma\delta_j, \tag{29}$$

and we associate this model with its parent normal model

$$y_{ij} \;=\; \alpha+\beta x_{ij}+\sigma\delta_j+\epsilon_{ij} \tag{30}$$

with $\text{var}(\epsilon_{ij})=1$, and common design matrix $\mathbf{X}$, regression parameters (α,β) and variance $\sigma^2=\text{var}(\delta_j)$ for the two models. We denote $\mathbf{V}_N=\text{var}(y_{ij}^{(N)})$ in (29) and $\mathbf{V}_B$ the generalized variance matrix for (30). Thus the information about (α,β) in (29) is approximated by $(\mathbf{F}_B=)\,\mathbf{X}^\top\mathbf{V}_B^{-1}\mathbf{X}$ and is equal to $(\mathbf{F}_N=)\,\mathbf{X}^\top\mathbf{V}_N^{-1}\mathbf{X}$ in (30). We have

$$\begin{aligned}\mathbf{F}_B \;&=\; \mathbf{X}^\top\mathbf{W}\mathbf{X}-\sigma^2\sum_j\mathbf{X}_j^\top\mathbf{W}_j\mathbf{z}_j\,\mathbf{z}_j^\top\mathbf{W}_j\mathbf{X}_j/g_j\\ &=\; \begin{pmatrix} \sum_j\frac{w_{+j}}{g_j} & \sum_j\frac{\mathbf{z}_j^\top\mathbf{W}_j\mathbf{x}_j}{g_j}\\ \sum_j\frac{\mathbf{x}_j^\top\mathbf{W}_j\mathbf{z}_j}{g_j} & \mathbf{x}^\top\mathbf{W}\mathbf{x}-\sum_j\frac{(\mathbf{x}_j^\top\mathbf{w}_j)^2}{g_j}\end{pmatrix},\end{aligned} \tag{31}$$

where $\mathbf{x}_j$ $(\mathbf{x})$ is the second column of $\mathbf{X}_j$ $(\mathbf{X})$, $\mathbf{z}_j$ is a vector of length n_j, and

$g_j = 1 + \sigma^2 w_{+j}$. In the corresponding formula for the normal model (30) the role of w_{+j} and $\mathbf{W}$ is taken by the number of observations in the cluster j (n_j), and the $n_j \times n_j$ identity matrix, respectively.

If we additionally assume that the average variance, ($c =$) w_{+j}/n_j, is constant within the clusters, then the information matrix $\mathbf{F}_N$ is equal to the information $\mathbf{F}_B$ for the dataset in which the design matrices $\mathbf{X}_j$ are replicated $1/c$ times in each cluster. Since $c \leq \frac{1}{4}$ we see that clustering has a much reduced impact on the bias of the estimators of the standard errors. This explains the apparent redundancy of the maximum likelihood methods over the GLS method in the interviewer variability example, and of the sample data analysis in the analysis of death rates in U.S. hospitals in Section 8.

The comparison of the logistic regression models with their parent models carries over to more complex patterns of variation; for instance, for the model

$$\text{logit}\{P(y_{ij} = 1|a_j, b_j)\} \;=\; a_j + b_j x_{ij} \tag{32}$$

with $(a_j, b_j) \sim \mathbf{N}_2(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ the approximate information matrix for $\boldsymbol{\beta}$ is equal to $\sum_j \mathbf{G}_j^{-1}\mathbf{X}_j^\top \mathbf{X}_j$, where $\mathbf{G}_j = \mathbf{I} + \mathbf{X}_j^\top \mathbf{W}\mathbf{X}_j \boldsymbol{\Sigma}$.

Information about the between-cluster variance σ^2 in the model (1) and its parent normal model permit a similar discussion. The information in the former is approximated by $\sum_j w_{+j}^2/(1 + \sigma^2 w_{+j})^2$, and in the normal case it is equal to $\sum_j n_j^2/(1 + \sigma^2 n_j)^2$. This implies that the binary models contain

relatively little information about σ^2, unless there are many clusters with large totals w_{+j}, and/or σ^2 is small. In the sense of the comparison above the information about σ^2 in (29) is comparable with the parent normal model (30) with n_j/w_{+j} times fewer observations in each cluster. Note that $n/(1+\sigma^2 n)$ is an increasing function of n, but its derivative is smaller than 1. Thus larger clusters contain more information about σ^2, but the increase is slower than when the same number of observations is added to the sample in new clusters.

8 Examples

8.1 Interviewer variability in an attitudinal survey

We illustrate the methods using a dataset kindly provided by Professor Shure from the Department of Psychology, UCLA. The data were collected in a survey of public awareness of political issues. The outcome variable is the respondent's perception of the government's role in his/her life, originally coded on the scale 1 – 5. For purposes of illustration we consider the dichotomous variable generated from this multinomial variable by recoding the values 1 and 2 into 0 and the values 3 – 5 into 1. There is one explanatory variable for the subjects (their gender, RSEX, coded 0 for females and 1 for males), and three variables defined for interviewers: their gender, ISEX, political opinion, IPOL, on the scale 1 – 4, from liberal to conservative, and (self-rated) concern for others, ICON, on the scale 1 – 3. Although the latter two variables

involve ordered categories, we will regard them throughout as quantitative variables, so as to reduce the number of estimated parameters and to simplify the discussion. The data contain records of 1008 respondents, each of them having been interviewed by one of the 40 interviewers. The workloads of the interviewers (numbers of respondents for each interviewer) vary between 12 and 85, although 32 interviewers have workloads smaller than 30 (see Table 1). Preliminary analysis as well as prior information imply that the between-cluster variance is likely to be very small, and possibly equal to zero, especially if differences in the interviewer-attributes are taken into account. This is largely confirmed by our analysis, although the size of the sampling variance of the estimate of σ^2 provides only weak evidence against large values of σ^2.

Results of model fits using the generalized least squares (GLS), the approximate maximum likelihood (AML), the restricted approximate maximum likelihood (RAML), the generalized estimating equations (GEE), the maximum likelihood using 3-, 5- and 9-point Gaussian quadrature (ML3, ML5 and ML9), and the restricted maximum likelihood using 9-point quadrature (REML) are displayed in Table 2.

The estimates and their estimated standard errors obtained by using 5-, 7- and 9-point quadrature differ by less than 10^{-4} (results for the 7-point quadrature are omitted), and for practical purposes it would suffice to use 3-point quadrature. The AML estimates and the standard errors for the

regression parameters are very close to the corresponding values for ML. The largest discrepancies, though still trivial, occur for GLS and for RAML. The discrepancies of the former are to be expected, in analogy with the (parent) normal case.

The estimate of the variance σ^2 in AML closely reproduces its ML counterpart, although the RAML and REML obtain substantially inflated estimates of the variance. The inflation factors are much higher than what could be expected by considering the number of regression parameters and the number of interviewers. The estimated standard error for the estimate of the standard deviation σ is greater than its AML counterpart but the estimated standard error for the estimates of the variance σ^2 is much smaller. This 'paradox' is due to the strong dependence of the information about σ and σ^2 on σ^2.

Figure 1 contains the plot of the profiles of the various approximations to the log-likelihood as functions of σ, and Figure 2 the same plots on the variance scale. It is preferable to derive naive confidence intervals based on standard errors for the variance, although in a wider range of values of σ^2 the dependence on the standard deviation σ appears to be much closer to a quadratic curve. Figures 1 and 2 also demonstrate that all the approximations to the log-likelihood, except ML3, are good for σ^2 in a wide range of realistic values.

The dependence of the restricted maximum likelihood deviances on the

variance closely resembles that for the full maximum likelihood. The correction term (28) varies insubstantially as a function of σ^2. The difference between AML (ML) and RAML (REML) estimates of σ is much smaller when fewer explanatory variables are used.

The GEE estimate of the 'working' correlation cannot be directly compared with the estimates of σ or σ^2, because it refers to a different scale. Note, however, that all the regression estimates are rather small, and so there is little variation in the fitted values $\mathbf{x}\hat{\boldsymbol{\beta}}$.

An approximate likelihood ratio test for the hypothesis of zero between-cluster variance can be carried out by comparing the values of –2log-likelihood (deviance), or their approximations, for the models with estimated σ^2 and the corresponding generalized linear model (assuming $\sigma^2 = 0$). In our case the difference of the deviances is approximately 0.10 for all methods, except RAML and REML, where it is equal to 0.66. Note that the RAML and REML deviances for the submodel with $\sigma^2 = 0$ is adjusted by the term (24); the RAML deviances cannot be compared with any deviances that refer to full maximum likelihood.

The iterations were terminated when the norm of the correction for the estimated parameters was smaller than 10^{-4} and the change in the value of the approximation to the log-likelihood (6) was smaller than 10^{-3}. Each method required three or four iterations, although all the methods except GLS use the GLS model fit as the starting solution. The times given in

Table 2 are the physical times (in seconds) that elapsed while fitting the model by the respective methods. The times for all the methods apart from GLS include the time taken to fit the GLS as the starting solution. Thus an iteration of AML, RAML, or GEE took 1.7 – 2.5 seconds. The time required for fitting ML increases with the number of quadrature points, about 10.5 seconds per point. All the model fits were carried out on an IBM/PC using the GAUSS software.

There appear to be consistent differences among male and female interviewees (the corresponding t-ratios are about 1.8), but the interviewer's attributes are not significant. Note however, that the estimated regression parameters, if taken at face value, are quite large: For a given respondent the logit of his/her response could differ by as much as 0.5 for a pair of interviewers with different sexes and extreme values of the attribute ICON. Even though there are 40 interviewers, the substantive conclusion of the analysis is that the data contain little information about interviewer variability (and even less information about the various pairwise comparisons of the interviewers). The estimates of the corresponding standard deviation σ are about 0.1, but even the value of 0.25 is quite feasible. Differences associated with such variation can be easily translated to the probability scale, and are in the present context substantial.

8.2 Death rates of Medicare patients at U.S. hospitals

The Health Care Finance Administration (HCFA) publishes annual reports giving for each acute care hospital in the U.S.A. the number of patients treated during the year and the number of deaths among those over 65 years insured by the Medicare system. The data are given for each of 14 diagnostic categories. The within-hospital death rates are compared with the national death rate for each diagnostic category and statistically significant comparisons (at a nominal 5% level) pointed out.

Clearly such a system of monitoring the quality of health care has a number of problems, including the choice of the outcome (died within 30 days of admittance, or survived), but one addressable issue is that of the risk associated with a patient's (hypothetical) selection of the hospital. Adjustment for the health condition of the patient at the time of admittance appears to be essential, but the relevant data, which consists of various measures of severity of the condition, is very costly to obtain because it requires time consuming abstracting from medical records by qualified staff.

As part of a large study assessing the quality of care under the Medicare Prospective Payment System (Kahn et al., 1990), a stratified probability sample of 297 hospitals was selected from the list of all U.S. acute care general hospitals active during the years 1981/82 and 1985/86. A set of standardized variables measuring patient's severity of condition at admission was extracted from a small number of randomly selected patient records (3 – 4

patients from each hospital and each time period per disease). An assessment of the quality of the medical care given to each patient was also performed. These assessments were coded in a quantitative variable called PROCESS. The age of each patient was also considered as a predictor variable.

For the complete national data for fiscal year 1986, analyzed by Jencks et al. (1988) using an alternative method, we consider the logistic random-effects analysis of variance model

$$\text{logit}\{P(y_{ij} = 1 \mid \delta_j)\} = \mu + \sigma\delta_j. \tag{33}$$

The estimates and standard errors for the parameters μ and σ^2 for four diagnostic conditions with high mortality rates are given in Table 3. We see that the data contain abundant information about between-hospital variation; each estimated variance is highly significant (using the t- or the likelihood ratio test). The estimated variances are substantial; for example, a pneumonia patient has an estimated probability of survival $1/\{1+\exp(-1.57-0.28)\} = 0.864$ in a hospital with $\delta_j = -1$, and $1/\{1+\exp(-1.57+0.28)\} = 0.784$ in a hospital with $\delta_j = +1$.

Unlike for the previous example, the estimate of the standard error of $\hat{\mu}$ is affected by the between-cluster variance quite substantially; the ratios of the standard errors for ML and for GLS are in the range $1.3 - 1.5$. This is a consequence of the data containing a large number of large clusters (several

hospitals admit annually more than 1,000 patients for a specific condition).

Adjustment for the explanatory variables would be expected to reduce the amount of between-hospital variation. We discuss here only the analysis for pneumonia. The death rate, not adjusted for severity, of the patients selected into the 1981/82 sample was 0.148 (1216 patients with mortality data), and the estimate of the variance σ^2 corresponding to the random effects ANOVA model (33) is 0.0960 (standard error 0.1610). In contrast, the death rate for the 1985/86 sample is 0.171 (1320 patients with mortality data), but the fitted variance is negative, equal to -0.0030 (standard error 0.1240).

Results for some of the logistic regressions for the survey data are displayed in Tables 4 (1981/82) and 5 (1985/86). Patient's severity is represented by the 11 variables used in Jencks et al. (1988), of which the variable APACHE II (Knaus, Draper and Wagner, 1985), is the most important. There are five stratifying variables (including dummy variables that categorize hospitals according to size), and the variable PROCESS is also considered. The variables APACHE II and PROCESS have been linearly rescaled to have mean zero and standard deviation 1. The ML estimates for the between-hospital variance are negative (zero) for most models for the 1985/86 data, whereas for 1981/82 they are positive, with exception of the model with adjustment for all the variables. Adjustment for severity appears to decrease the estimate of the between-cluster variance, while adjustment for process has the opposite effect, though to a much lesser degree. However,

the estimated standard errors associated with these estimates of σ^2 are so large that we have little confidence that severity adjustment does reduce the differences among the hospitals. The data with all the severity measures have too little information about between-cluster variation. The negative estimated variances are clearly unrealistic; in fact they could not be realized in hospitals with more than about 100 patients.

It would seem that larger samples are required to obtain a meaningful estimate for σ^2. The approximate information for σ^2 given by (21) can be used to decide about the survey design that would improve or optimize the information about σ^2. Suppose, for simplicity, that the same number of patients, n, is to be sampled from each selected hospital, and a total of N patients will be selected to the survey. The approximate information for σ^2 is equal to $\frac{1}{2}Nnw^2/(1+nw\sigma^2)$, where $w = p(1-p)$ is the common conditional variance (given $\delta_j = 0$), and for fixed N, w and σ^2 it has a unique maximum for $n^* = 1/w\sigma^2$. For pneumonia we have $w \approx 0.15$ and almost certainly $\sigma^2 < 0.3$, and so it is very likely that $n^* > 20$. This suggests that a design with fewer hospitals and more patients from each selected hospital would be much more informative about σ^2. However, design with independent observations (one patient per hospital) is most informative about the regression parameters. A suitable tradeoff is likely to be closer to the design with fewer sampled hospitals.

9 Simulations

The methods discussed in Sections 3 and 4 were compared in a simulation study. Data were generated according to the simple logistic regression model with a random intercept:

$$\text{logit}\{P(y_{ij} = 1 \mid \delta_j)\} = \alpha + \beta x_{ij} + \sigma\delta_j, \tag{34}$$

with 40 clusters (j), four of them containing $21, 22, \ldots, 30$ observations each. The regression parameters were set to $\alpha = 0$ and $\beta = 1$. The values of the regressor x were drawn from the uniform distributions on $(-1, 1)$ for one set of 100 simulations, and on $(1, 3)$ for another set of 100 simulations. The (regression) designs are referred to as U(−1,1) and U(1,3). The values of the standard deviation σ were set to $0, 0.1, \ldots, 1$ in each a set of simulations.

The purpose of the simulations was to compare the estimators of the parameters α, β, and σ (σ^2), and the standard errors of these estimators, to assess the importance of taking account of between-cluster variation in estimating α and β, as well as to compare the computational complexity of the methods. Of interest is the agreement of the AML, RAML and GEE estimators with their exact ML counterparts and the accuracy of the estimated standard errors for predicting the observed mean squared errors.

We summarize results separately for each estimated quantity (parameter or standard error):

1. *Slope β:* The biases of the GLS, AML, RAML, ML9 and REML9 estimators of the slope β are plotted in Figures 3 and 4 for the respective designs U(−1,1) and U(1,3). The AML, RAML and GLS estimators of the slope have nearly identical means, and their biases are essentially a decreasing function of the variance σ^2. The GEE estimator is essentially indistinguishable from the AML estimator, and is therefore omitted from the plots. The ML9 and REML9 estimators are also nearly identical, but their biases are much smaller than those of the GLS, AML and RAML estimators. However, the difference in the biases is substantial only for $\sigma \geq 0.4$ for U(−1,1), and $\sigma \geq 0.7$ for U(1,3). It is rather counterintuitive that the bias in much smaller in the U(1,3) design which contains much less information about the slope.

2. *Variance σ^2:* The means of the AML, RAML, ML9 and REML9 estimators of the variance σ^2 are plotted in Figures 5 and 6 for the respective designs U(−1,1) and U(1,3). The bias of the REML9 estimator is negligible for up to $\sigma = 0.9$. The ML9 and REML9 have comparable biases, and the bias of the AML estimator is the largest. However, the bias of the latter is still ignorable for $\sigma^2 \leq 0.4$ in both designs. Again the bias of all four estimators is much smaller for the U(1,3) design which contains less information about σ^2. Note that in ML and REML methods the standard deviations are estimated, and so the corresponding estimator of the variance is nonnegative. As a result these estimators of the variance have a positive bias for small values of σ^2.

3. *Mean squared error of the estimators of the slope:* Figures 7 and 8 contain the plots of the mean squared errors for the five studied estimators. The GLS estimator is only slightly less efficient than its AML and RAML counterparts for the U(−1,1) design, and its efficiency breaks down only for $\sigma > 0.5$ in the U(1,3) design. The ML9 and REML9 estimators of β are substantially more efficient for large values of σ in U(−1,1) design (the efficiency reaches about 135% for $\sigma = 1$), but they are less efficient than the AML and RAML estimators for the U(1,3) design throughout the range of σ. Note that for the U(−1,1) design the mean squared error is an increasing function of σ for $\sigma > 0.3$, but in the U(1,3) no such trend can be observed.

4. *Standard errors for the estimators of the variance* σ^2 : AML and RAML methods estimate the variance σ^2, and in our implementation negative estimates of the variance were allowed. These methods can be adapted for estimation of the standard deviation σ in the obvious way, and the corresponding standard error can be calculated by application of the chain rule. However, the sign of the value of the estimate is not determined, and therefore the definition of the corresponding mean squared error is ambiguous. The ML9 and REML9 methods estimate the standard deviation σ, from which an estimator of σ^2 can be defined, but it does not allow negative values of the variance. The sampling distribution of the estimators of σ would be expected to be a censored normal. The histogram of the 48 positive estimates of σ by AML method in U(−1,1) design (Figure 9) provides strong evidence to

the contrary. Substantially smaller proportion of the estimates of σ are close to 0 than would be expected. Generally, the distributions of the variance estimators appear to resemble the normal distribution more closely than the estimators of standard deviation, although for large values of the variance the difference is not substantial. We therefore compare the ML9 and REML9 estimators for the standard errors with their AML and RAML counterparts only for $\sigma \geq 0.5$, where no negative estimates of σ^2 have occurred. Figures 10 and 11 contain plots of the means of the four estimators of standard errors. The correction for bias appears to increase the standard errors, though only marginally. The approximate methods, AML and RAML, yield more efficient estimators than their exact counterparts, ML9 and REML9, though the difference is insubstantial.

We have replicated a small sample of the simulations of the ML9 estimator using 7- and 11- point Gaussian quadrature. No corresponding sets of estimated quantities (parameters and standard errors) differed by more than 10^{-4}.

The value of the deviance (−2 log-likelihood) was obtained for each method. It provides information about the quality and power of the (approximate) likelihood ratio tests for the hypothesis of independence ($\sigma^2 = 0$). The sampling distributions for the corresponding deviance differences for all four methods closely resemble the χ^2_1 distribution. Figure 12 contains the corresponding qq-plot for the AML method, U(−1,1) design.

The naive t-ratio test for zero variance closely agrees with the likelihood ratio criterion. The power of these tests is very low; the observed power of the likelihood ratio test for the null hypothesis of $\sigma^2 = 0$ when the true variance is, say, 0.09, is only 44% for the U(−1,1) design, and 19% for the U(1,3) design (at the 5% level of significance). Clearly, with such designs there is little scope for modelling more complex covariance structures than the equicovariance one.

10 Discussion

The approximation to the log-likelihood and its partial derivatives for logistic regression with random coefficients provides an alternative derivation of the GEE approach of Liang and Zeger (1986). The approximation highlights the problematic nature of both methods of estimation when the estimated between-cluster variation is large.

The loss of information about the regression parameters attributable to clustering in logistic regression is much smaller than in the normal regression, and, in addition to the cluster sizes it depends on the distribution of the predicted variances w_{ij}.

The AML approach discussed in this paper is a model- and estimation-framework parallel with the GEE approach of Liang and Zeger (1986). In addition, the AML approach is supported with an approximation to the log-likelihood, and for random regression coefficients, with a model descrip-

tion which enables data simulation, and in principle, detailed model diagnostic procedures, such as those described by Pregibon (1981). Also the parametrization implied by these models is a natural one, and is easy to transform from the linear scale to the scale of the outcomes, such as to the probabilities in logistic regression.

References

Anderson, D. and Aitkin, M. (1985) Variance component models with binary response: Interviewer variability. *J. R. Statist Soc. B* **47**, 203–210.

Bonney, G.E. (1987) Logisitic regression for dependent binary observations. *Biometrics* **43**, 951–973.

Breslow, N. (1984) Extra-Poisson variation in log-linear models. *Applied Statistics* **33**, 38–44.

Connolly, M. and Liang, K.Y. (1988) Conditional logistic regression models for correlated binary data. *Biometrika* **75**, 501–506.

Dempster, A.P., Laird N.M. and Rubin D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B* **39**, 1–38.

Goldstein, H. (1986) Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika* **73**, 43–56.

Hartley, H.O. and Rao, J.N.K. (1967) Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika* **54**, 93–108.

Harville, D.A. (1974) Bayesian inference for variance components using only error contrasts. *Biometrika* **61**, 383–385.

Harville, D.A. (1977) Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Ass.* **72**, 320–340.

Harville, D.A. and Mee, R.W. (1984) A mixed model procedure for analyzing ordered categorical data. *Biometrics* **40**, 393–408.

Holt, D. and Scott, A.J. (1982) Regression analysis using survey data. *The Statistician* **30**, 169–177.

Im, S. and Gianola, D. (1988) Mixed models for binomial data with an application to lamb mortality. *Applied Statistics* **37**, 196–204.

Jencks, S., F., Daley, J., Draper, D., Thomas, N., Lenhart, G., and Walker, J. (1988) Interpreting hospital mortality data: The role of clinical risk adjustment. *Journal of the American Medical Association* **260**, 3611–3616.

Jennrich, R.I. and Schluchter, M.D. (1986) Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* **42**, 805–820.

Kahn, K., Rubenstein, L.V., Draper, D., Kosecoff, J., Rogers, W.H., Keeler, E.B., and Brook, R.H. (1990) Effects of the DRG-based prospective payment system on quality of care for hospitalized Medicare patients: An introduction to the series. *Journal of the American Medical Association*, **264**, 1953-1955.

Knaus, W.A., Draper, E.A., and Wagner, D.P. (1985) APACHE II: A severity of disease classification system for severely ill patients. *Crit. Care Med.* **13**, 818–829.

Laird, N.M. and Ware, J.H. (1982) Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

Lange, N. and Laird, N.M. (1989) The effect of covariance structure on variance estimation in balanced growth-curve models with random parameters. *J. Amer. Statist. Ass.* **84**, 241–247.

Liang K.Y. and Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Lindstrom, M.J. and Bates, D.M. (1988) Newton-Raphson and EM algorithms for linear mixed-effects models for repeated measures data. *J. Amer. Statist. Assoc.* **83**, 1014–1022.

Longford, N.T. (1987) A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika* **74**, 812–827.

Longford, N.T. (1988) A quasilikelihood adaptation for variance component analysis. *Proceedings of the Section on Statistical Computing of the American Statistical Association.* Alexandria VA.

McCullagh, P. and Nelder, J.A. (1990) *Generalized Linear Models.* Monographs on Statistics and Applied Probability, Chapman and Hall, 2nd edition. London New York.

Nelder, J.A. and Pregibon, D. (1987) An extended quasi-likelihood function. *Biometrika* **74**, 221–232.

Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized linear models. *J. R. Statist. Soc. A* **135**, 370–84.

Patterson, H.D. and Thompson, R. (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–554.

Pregibon, D. (1981) Logistic regression diagnostics. *Ann. Statist.* **9**, 705–724.

Prentice, R.L. (1986) Binary regression using an extended beta-binomial distribution with discussion of correlation induced by covariate measurement errors. *J. Amer. Statist. Ass.* **81**, 321–327.

Prentice, R.L. (1988) Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**, 1033–1048.

Qu, Y.S., Williams, G.W. Beck, G.J. and Goormastic, M. (1987) A generalized model of logistic regression for correlated data. *Communications in Statistics - Theory and Methods* **16**, 3447–3476.

Rosner, B. (1984) Multivariate methods in ophtalmology with application to other paired data situations. *Biometrics* **40**, 1025–1035.

Rosner, B. (1989) Multivariate methods for clustered binary data with more than one level of nesting. *J. Amer. Statist. Ass.* **84**, 373–80.

Searle, S.R. (1971) Topics in variance component estimation. *Biometrics* **27**, 1–76.

Stanek, E.J. III and Diehl, S.R. (1988) Growth curve models of repeated binary response. *Biometrics* **44**, 973–983.

Stiratelli, R., Laird, N. and Ware, J.H. (1984) Random-effect models for serial observations with binary response. *Biometrics* **40**, 961–971.

Stokes, L. (1988) Estimation of interviewer effects for categorical items in a random digit dial telephone survey. *J. Amer. Statist. Ass.* **83**, 623–630.

Stram, D.O., Wei, L.J. and Ware, J.H. (1988) Analysis of repeated order categorical outcomes with possibly missing observations and time dependent covariates. *J. Amer. Statist. Ass.* **83**, 631–637.

Thompson, R. and Meyer, K. (1986) Estimation of variance components: What is missing in the EM algorithm? *J. Statist. Comput. Simul.* **24**, 215–230.

Williams, D.A. (1982) Extra-binomial variation in logistic linear models. *Applied Statistics* **31**, 144–148.

Zeger, S.L. and Liang, K. Y. (1986) Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121–130.

Zeger, S.L., Liang, K.Y. and Albert, P.S. (1988) Models for longitudinal data: A generalized estimating equation approach. *Biometrics* **44**, 1049–1060.

Zeger, S.L. and Qadish, B. (1988) Markov regression models for time series: quasilikelihood approach. *Biometrics* **44**, 1019–1031.

Zhao, L.P. and Prentice, R.L. (1990) Correlated binary regression using a quadratic exponential model. *Biometrika* **77**, 642–648.

Appendix
Approximation for the likelihood of the random effects model with non-normally distributed outcomes

We consider the density of a distribution in the exponential class

$$f(y,\theta,\tau) \;=\; \exp[a(\tau)\{y\theta - b(\theta)\} + c(y;\tau)], \tag{1}$$

in which the parameter θ is a function of the linear predictor $\mathbf{x}\boldsymbol{\beta}$. We will use the following expansion:

$$y\theta(\mathbf{x}\boldsymbol{\beta}+\mathbf{z}\boldsymbol{\delta}) - b\{\theta(\mathbf{x}\boldsymbol{\beta}+\mathbf{z}\boldsymbol{\delta})\}$$

$$\begin{aligned}
&\approx \quad y\theta(\mathbf{x}\boldsymbol{\beta}) - b\{\theta(\mathbf{x}\boldsymbol{\beta})\} + \mathbf{z}\boldsymbol{\delta}[y\theta'(\mathbf{x}\boldsymbol{\beta}) - b'\{\theta(\mathbf{x}\boldsymbol{\beta})\}\theta'(\mathbf{x}\boldsymbol{\beta})] \\
&+ \quad \tfrac{1}{2}(\mathbf{z}\boldsymbol{\delta})^2[y\theta''(\mathbf{x}\boldsymbol{\beta}) - b''\{\theta(\mathbf{x}\boldsymbol{\beta})\}\{\theta(\mathbf{x}\boldsymbol{\beta})\}^2 - b'\{\theta(\mathbf{x}\boldsymbol{\beta})\}\theta''(\mathbf{x}\boldsymbol{\beta})] \\
&= \quad A_0(\mathbf{x}\boldsymbol{\beta}) + \mathbf{z}\boldsymbol{\delta}A_1(\mathbf{x}\boldsymbol{\beta}) - \tfrac{1}{2}(\mathbf{z}\boldsymbol{\delta})^2 A_2(\mathbf{x}\boldsymbol{\beta}).
\end{aligned} \tag{2}$$

We assume that $a(\tau) > 0$ and $A_2(\theta) > 0$ for all τ and θ. Note that these functions are related to the variance of the distribution (1).

Suppose each observation $i = 1, \ldots, n$ of a cluster has the density $f(y_i, \theta_i, \tau)$ with $\theta_i = \theta(\mathbf{x}_i\boldsymbol{\beta} + \mathbf{z}_i\boldsymbol{\delta})$, where $\mathbf{x}_i$ and $\mathbf{z}_i$ are given vectors and $\boldsymbol{\delta} \sim \mathbf{N}_r(\mathbf{0}, \boldsymbol{\Sigma}_1)$ for a positive definite matrix $\boldsymbol{\Sigma}_1$. The observations are assumed conditionally independent given $\boldsymbol{\delta}$. The joint density for the cluster can be approximated, using the expansion (2), as

$$\{(2\pi)^r \det \boldsymbol{\Sigma}_1\}^{-\frac{1}{2}} \int_{\mathbf{R}^r} \cdots \int \prod_{i=1}^{n} f\{y_i, \theta(\mathbf{x}_i\boldsymbol{\beta} + \mathbf{z}_i\boldsymbol{\delta}), \tau\} \exp\left(-\frac{1}{2}\boldsymbol{\delta}^\top \boldsymbol{\Sigma}_1 \boldsymbol{\delta}\right) d\boldsymbol{\delta}$$

$$\approx \{(2\pi)^r \det \boldsymbol{\Sigma}_1\}^{-\frac{1}{2}} \textstyle\prod_{i=1}^{n} f\{y_i, \theta(\mathbf{x}_i\boldsymbol{\beta}), \tau\}$$

$$\int \cdots \int \exp\left[a(\tau)\left\{\mathbf{A}_1\mathbf{Z}\boldsymbol{\delta} - \tfrac{1}{2}\boldsymbol{\delta}^\top(\mathbf{Z}^\top\mathbf{A}_2\mathbf{Z} + \boldsymbol{\Sigma}^{-1})\boldsymbol{\delta}\right\}\right] d\boldsymbol{\delta}$$

$$= \textstyle\prod_{i=1}^{n} f\{y_i, \theta(\mathbf{x}_i\boldsymbol{\beta}), \tau\}[\{a(\tau)\}^n \det \mathbf{G}]^{-\frac{1}{2}} \exp\{\tfrac{1}{2}a(\tau)\mathbf{e}^\top\mathbf{e} - \tfrac{1}{2}\mathbf{e}^\top\mathbf{V}_1^{-1}\mathbf{e}\}, \tag{3}$$

where $\mathbf{e} = \mathbf{A}_1\mathbf{A}_2^{-1}$, $\mathbf{V}_1 = a^{-1}(\tau)(\mathbf{A}_2 + \mathbf{A}_2\mathbf{Z}\boldsymbol{\Sigma}_1\mathbf{Z}^\top\mathbf{A}_2)$, $\mathbf{G} = \mathbf{I}_r + \mathbf{Z}^\top\mathbf{A}_2\mathbf{Z}\boldsymbol{\Sigma}_1$, $\mathbf{A}_1 = \{A_1(\mathbf{x}_1\boldsymbol{\beta}), \ldots, A_1(\mathbf{x}_n\boldsymbol{\beta})\}$, $\mathbf{A}_2 = \mathrm{diag}\{A_2(\mathbf{x}_1\boldsymbol{\beta}), \ldots, A_2(\mathbf{x}_n\boldsymbol{\beta})\}$, and $\mathbf{Z} = (\mathbf{z}_1^\top, \ldots, \mathbf{z}_n^\top)^\top$. Note that

$$\mathbf{V}_1^{-1} = a(\tau)(\mathbf{A}_2^{-1} - \mathbf{Z}\boldsymbol{\Sigma}_1\mathbf{G}^{-1}\mathbf{Z}^\top).$$

Next suppose there are clusters $j = 1, \ldots, N_2$ with n_j observations each, with vectors of outcomes y_j and design matrices $\mathbf{X}_j = (\mathbf{x}_{1j}^\top, \ldots, x_{n_j j}^\top)^\top$ and $\mathbf{Z}_j$, and conditionally on $\boldsymbol{\gamma} \in \mathbf{R}^s$ each cluster has the (approximated) joint density (3) with $\theta(\mathbf{x}_i\boldsymbol{\beta})$ replaced by $\theta(\mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{u}_{ij}\boldsymbol{\gamma})$, and let $\mathbf{U}_j$ be the $n_j \times s$ matrix containing the rows $\mathbf{u}_{ij}$. Suppose these clusters are conditionally independent and that $\boldsymbol{\gamma} \sim \mathbf{N}_s(\mathbf{0}, \boldsymbol{\Sigma}_2)$, where $\boldsymbol{\Sigma}_2$ is a positive definite matrix. The joint density for these $N = \sum_{j=1}^{N_2} n_j$ observations can be approximated, using (2), as

$$F\{\mathbf{Y}, \boldsymbol{\Theta}(\mathbf{X}\boldsymbol{\beta}), \tau\} \left[\{a(\tau)\}^N \prod_{j=1}^{N_2} \det \mathbf{G}_j\right]^{-\frac{1}{2}}$$

$$\{(2\pi)^s \det \boldsymbol{\Sigma}_2\}^{-\frac{1}{2}} \exp\left\{\tfrac{1}{2}a(\tau) \textstyle\sum_j \mathbf{e}_j^\top \mathbf{e}_j - \tfrac{1}{2} \sum_j \mathbf{e}_j^\top \mathbf{V}_{1j}^{-1} \mathbf{e}_j\right\}$$

$$\int \ldots \int \exp\left[a(\tau)\left\{\textstyle\sum_j \mathbf{A}_{1j}\mathbf{U}_j\boldsymbol{\gamma} - \tfrac{1}{2}\boldsymbol{\gamma}^\top\left(\sum_j \mathbf{U}_j^\top \mathbf{A}_{2j}\mathbf{U}_j + \boldsymbol{\Sigma}_2^{-1}\right)\boldsymbol{\gamma}\right\}\right] d\boldsymbol{\gamma}$$

$$= F\{\mathbf{Y}, \boldsymbol{\Theta}(\mathbf{X}\boldsymbol{\beta}), \tau\} \left[\{a(\tau)\}^N \prod_{j=1}^{N_2} \det \mathbf{G}_j \det \mathbf{H}\right]^{-\frac{1}{2}} \exp\left\{\tfrac{1}{2}a(\tau)\mathbf{d}^\top\mathbf{d} - \tfrac{1}{2}\mathbf{d}^\top\mathbf{V}_2^{-1}\mathbf{d}\right\}, \tag{4}$$

where

$$F\{\mathbf{Y}, \boldsymbol{\Theta}(\mathbf{X}\boldsymbol{\beta}), \tau\} = \prod_{j}^{N_2} \prod_{i}^{n_j} f\{y_{ij}, \theta(\mathbf{x}_{ij}\boldsymbol{\beta}), \tau\},$$

$\mathbf{H} = \mathbf{I} + \sum_j \mathbf{U}_j^\top \mathbf{B}_2^\top \mathbf{V}_{1j}^{-1} \mathbf{B}_2 \mathbf{U}_j \boldsymbol{\Sigma}_2$, $\mathbf{d} = (\mathbf{e}_1^\top, \ldots, \mathbf{e}_{N_2}^\top)^\top$, $\mathbf{e}_j, \mathbf{G}_j, \mathbf{A}_{1j}, \mathbf{A}_{2j}$, and $\mathbf{V}_{1j}$ are defined as the corresponding vectors or matrices $\mathbf{e}, \mathbf{G}, \mathbf{A}_1, \mathbf{A}_2$, and $\mathbf{V}_1$ for cluster j, $\mathbf{B}_1 = (\mathbf{A}_{11}^\top, \mathbf{A}_{12}^\top, \ldots, \mathbf{A}_{1N_2}^\top)^\top$, $\mathbf{B}_2 = \mathrm{diag}(\mathbf{A}_{21}^\top, \mathbf{A}_{22}^\top, \ldots, \mathbf{A}_{2N_2}^\top)^\top$, and

$$\mathbf{V}_2 \;=\; \mathrm{diag}(\mathbf{V}_{1j}) + a^{-1}(\tau)\mathbf{B}_2\mathbf{U}\boldsymbol{\Sigma}_2\mathbf{U}^\top\mathbf{B}_2^\top,$$

where $\mathbf{U} = (\mathbf{U}_1^\top, \ldots, \mathbf{U}_{N_2}^\top)^\top$.

Extensions to further layers of nesting are analogous; in the appropriately altered notation we have the following approximation for the joint density of a cluster of order k (containing N_{k-1} clusters of order $k-1$, each of these containing clusters of order $k-2$, and so on):

$$F\{\mathbf{Y}, \boldsymbol{\Theta}(\mathbf{X}\boldsymbol{\beta}), \tau\} \left[a(\tau)\right]^N \prod_{h=1}^{k} \prod_{i=1}^{N_h} \det \mathbf{G}_{hi} \Bigg]^{-\frac{1}{2}} \exp\left(\frac{1}{2}\mathbf{e}^\top\mathbf{e} - \frac{1}{2}\mathbf{e}^\top\mathbf{V}_k^{-1}\mathbf{e}\right), \quad (5)$$

where the (generalized variance) matrix $\mathbf{V}_k$ is defined recursively:

$$\mathbf{V}_k \;=\; \mathrm{diag}(\mathbf{V}_{k-1j}) + \mathbf{A}_2\mathbf{Z}_k\boldsymbol{\Sigma}_k\mathbf{Z}_k^\top\mathbf{A}_2,$$

and $\mathbf{G}_k = \mathbf{I} + \sum_j \mathbf{Z}_{kj}\mathbf{V}_{k-1}^{-1}\mathbf{Z}_{kj}^{\top}\boldsymbol{\Sigma}_k$. Note that the product of the determinants in (5) is equal to $\det\mathbf{V}$.

An approximate maximum likelihood estimator for all the unknown parameters can be defined as the maximizer of (5). The first- and second-order partial derivatives of the logarithm of (5) with respect to the regression parameters $\boldsymbol{\beta}$ have the approximations (17) and (18) (the dependence of the weights on the linear predictor has to be ignored). This implies the generalized least squares estimator. The first-order partial derivative with respect to a covariance structure parameter θ is

$$\frac{\partial l}{\partial \theta} = \frac{1}{2}\left\{-\mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta}\right) + \mathbf{e}^{\top}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta}\mathbf{V}^{-1}\mathbf{e}\right\} \tag{6}$$

and the expectation of the second-order partial derivative has the approximation (setting $\mathbf{E}\,(\mathbf{e}\mathbf{e}^{\top}) \approx \mathbf{V}$)

$$-\mathbf{E}\left(\frac{\partial^2 l}{\partial \theta_1 \theta_2}\right) = \frac{1}{2}\mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_1}\,\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_2}\right). \tag{7}$$

The approximations (3) – (7) can be extended for singular matrices $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$, ..., since the approximate joint density (5) can be analytically extended to the boundary of the parameter space for these variance matrices. Also,

extension for non-constant scale $a(\tau)$ is trivial.

Longford (1988) gives an approximation to the joint density for clustered observations, similar to the derivation presented here, with the conditional density (1) replaced by the extended quasilikelihood function of Nelder and Pregibon (1987).

Table 1: The workload of the interviewers; Interviewer Variability Example

12	12	13	14	15	15	16	17	17	17
18	18	18	19	19	20	20	20	21	21
22	23	23	23	23	24	24	24	24	26
26	28	30	31	36	42	45	47	60	85

Table 2. Regression model fits to the interviewer data.

The methods are: generalized least squares (GLS), approximate maximum likelihood (AML), restricted approximate maximum likelihood (RAML), generalized estimating equations (GEE), exact maximum likelihood using 3-, 5- and 9- point quadrature (ML3, ML5, and ML9, respectively), and restricted exact maximum likelihood (REML), using 9-point quadrature. The estimates are given in the columns 2 – 7 and the corresponding standard erors underneath in parentheses. The deviances are the values of -2 log-likelihood.

Legend:

[1] The RAML deviance for the GLS model fit

[2] Estimate of the working correlation (GEE)

Table 2: Regression model fits to the interviewer data.

	PARAMETER ESTIMATES							MODEL FITTING INFORMATION		
METHOD	Intercept	RSEX	IPOL	ICON	ISEX	σ^2	σ	Iterations	Deviance	Comp. Time
GLS	-.5965 (.2937)	-.2705 (.1515)	-.0708 (.0707)	-.1793 (.1370)	-.1142 (.1671)			4	1071.152	5.27
AML	-.5906 (.3041)	-.2686 (.1518)	-.0736 (.0729)	-.1823 (.1418)	-.1128 (.1740)	.0136 (.0486)	.1166 (.2085)	3	1071.055	10.33
RAML	-.5817 (.2937)	-.2656 (.1515)	-.0779 (.0707)	-.1875 (.1370)	-.1101 (.1671)	.0391 (.0524)	.1976 (.1327)	4	1092.459 1093.114[1]	13.07
GEE	-.5900 (.3380)	-.2680 (.1478)	-.0735 (.0743)	-.1827 (.1467)	-.1132 (.2051)	.0324[2]		3		12.85
ML3	-.5934 (.3050)	-.2692 (.1519)	-.0739 (.0520)	-.1825 (.1423)	-.1126 (.1741)		.1162 (.1993)	4	1071.056	59.59
ML5	-.5931 (.3052)	-.2692 (.1519)	-.0739 (.0521)	-.1826 (.1424)	-.1128 (.1741)		.1173 (.2017)	4	1071.055	80.68
ML9	-.5931 (.3052)	-.2692 (.1519)	-.0739 (.0521)	-.1826 (.1424)	-.1128 (.1741)		.1173 (.2016)	4	1071.055	122.75
REML	-.5841 (.2980)	.2661 (.1508)	-.0774 (.0717)	.1889 (.1386)	-.1130 (.1652)		.1953 (.1360)	4	1092.373 1093.024	84.22

Table 3: Analysis of the complete national data for the four most frequent conditions.
Standard errors are given in parentheses ( ) and the ML9 estimates in braces [ ]. The standard errors for the ML9 estimates are omitted (to conserve space); they differ from their counterparts by less than 0.0002 for the mean logit, and less than 0.0006 for the variance σ^2.

DEATH RATES AND THEIR BETWEEN-HOSPITAL VARIATION
COMPLETE NATIONAL DATA, FISCAL 1986

Conditions	Hospitals	Patients Died	Death rt (%)	GLM	AML [ML]	σ^2	Lklh. ratio
Pneumonia	5628	415,179		-1.4645	-1.5057	0.2830	1689.03
		77,961	18.78		[-1.4811]	[0.2846]	[1710.12]
				(0.0040)	(0.0059)	(0.0061)	
Heart Failure	5541	465,229		-1.7106	-1.7184	0.2109	559.45
		71,223	15.31		[-1.7037]	[0.2071]	[553.63]
				(0.0041)	(0.0054)	(0.0068)	
Stroke	5310	298,306		-1.3664	-0.13549	0.2598	813.81
		60,617	20.32		[-1.3370]	[0.2518]	[805.81]
				(0.0046)	(0.0063)	(0.0075)	
Heart Attack	5285	278,114		-1.0717	-1.0595	0.2085	507.21
		70,942	25.51		[-1.0406]	[0.2033]	[499.11]
				(0.0044)	(0.0057)	(0.0071)	

Table 4: Logistic regression for death rates (pneumonia), with adjustment for severity, stratification and PROCESS (as indicated in the rows), year 1981/82.
HOSP denotes the random effects due to the hospitals, STRAT the stratifying variables, PROC the process variable, and SEVER the measures of severity (they include APACHE as a variable). The standard errors for $\hat{\sigma}^2$ and given in parentheses ( ), and the deviance corresponding to $\sigma^2 = 0$ (the GLS deviance) in brackets [ ].

PARAMETER ESTIMATES (ST. ERRORS)					
Adjustment for	APACHE	PROCESS	$\hat{\sigma}^2$	Deviance [GLS deviance]	Regression parameters
HOSP			0.0960 (0.1610)	1019.27 [1019.66]	1
HOSP, STRAT			0.0711 (0.1579)	1014.63 [1014.85]	6
HOSP, PROC		-0.1424 (0.0840)	0.1099 (0.1622)	1016.31 [1016.82]	2
HOSP, SEVER	0.9556 (0.1074)		0.0036 (0.2123)	692.82 [692.82]	12
HOSP, SEVER PROC	0.9883 (0.1100)	-0.1540 (0.1013)	0.0092 (0.2130)	690.50 [690.51]	13
HOSP, SEVER PROC, STRAT	0.9841 (0.1107)	-0.1626 (0.1030)	-0.0093 (0.2104)	687.57 [687.57]	18

Table 5: Logistic regression for death rates (pneumonia), with adjustment for severity, stratification and PROCESS, 1985/86.
The same notation is used as in Table 4.

PARAMETER ESTIMATES (ST. ERRORS)					
Adjustment for	APACHE	PROCESS	$\hat{\sigma}^2$	Deviance [GLS deviance]	Regression parameters
HOSP			-0.0030 (0.1240)	1208.60 [1208.60]	1
HOSP, STRAT			-0.0289 (0.1207)	1204.81 [1204.86]	6
HOSP, PROC		-0.2721 (0.0762)	0.0172 (0.1263)	1195.78 [1195.80]	2
HOSP, SEVER	0.8032 (0.0937)		-0.1053 (0.1461)	915.31 [915.77]	12
HOSP, SEVER PROC	0.8459 (0.0960)	-0.1994 (0.0884)	-0.1172 (0.1456)	910.29 [910.84]	13
HOSP, SEVER PROC, STRAT	0.8521 (0.0963)	-0.1876 (0.0892)	-0.1542 (0.1498)	906.61 [907.53]	18

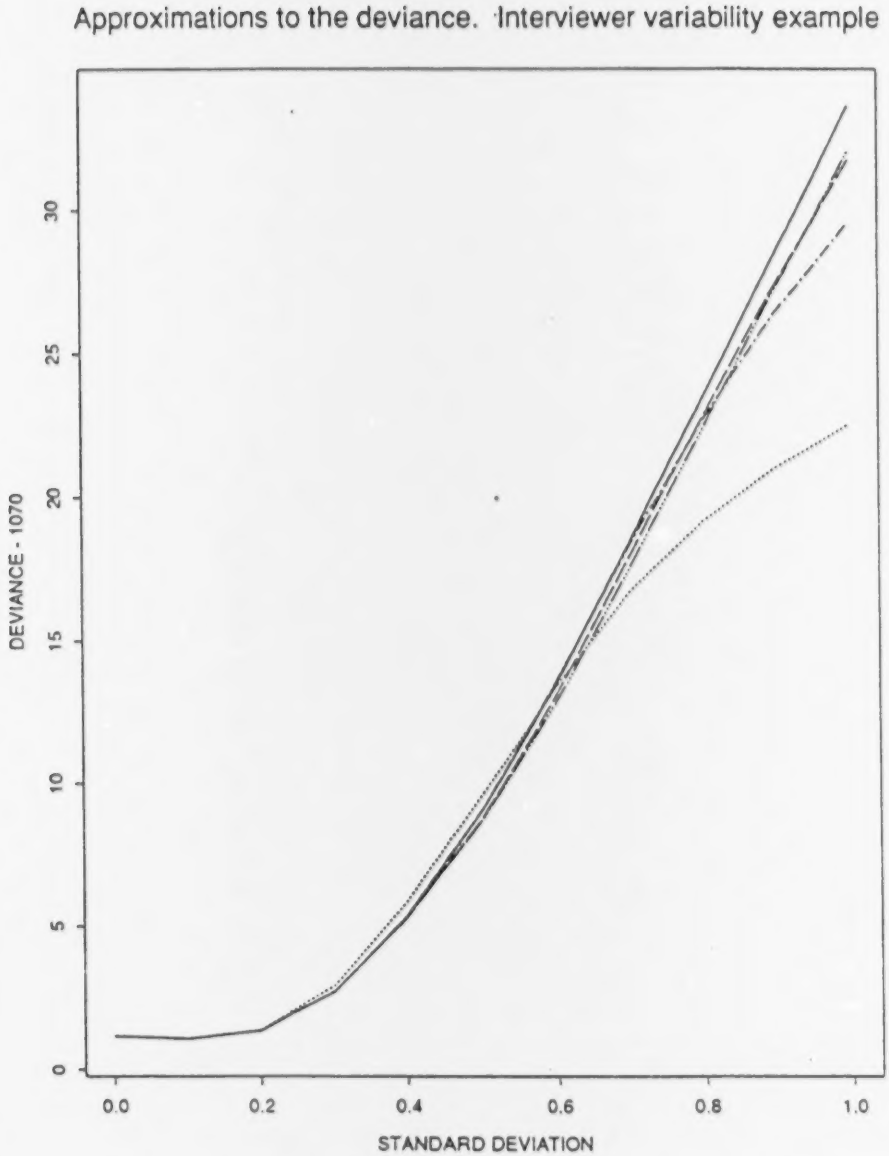

Figure 1: Profiles of the approximations to the −2 log-likelihood for the Interviewer variability data as a function of the standard deviation σ.
The methods of approximation are: ——— AML, approximate log-likelihood (16); ····· ML3; —·—· ML5; — — ML7; —·—· ML9 (MLk stands for the k-point Gaussian quadrature approximation to the −2 log-likelihood).

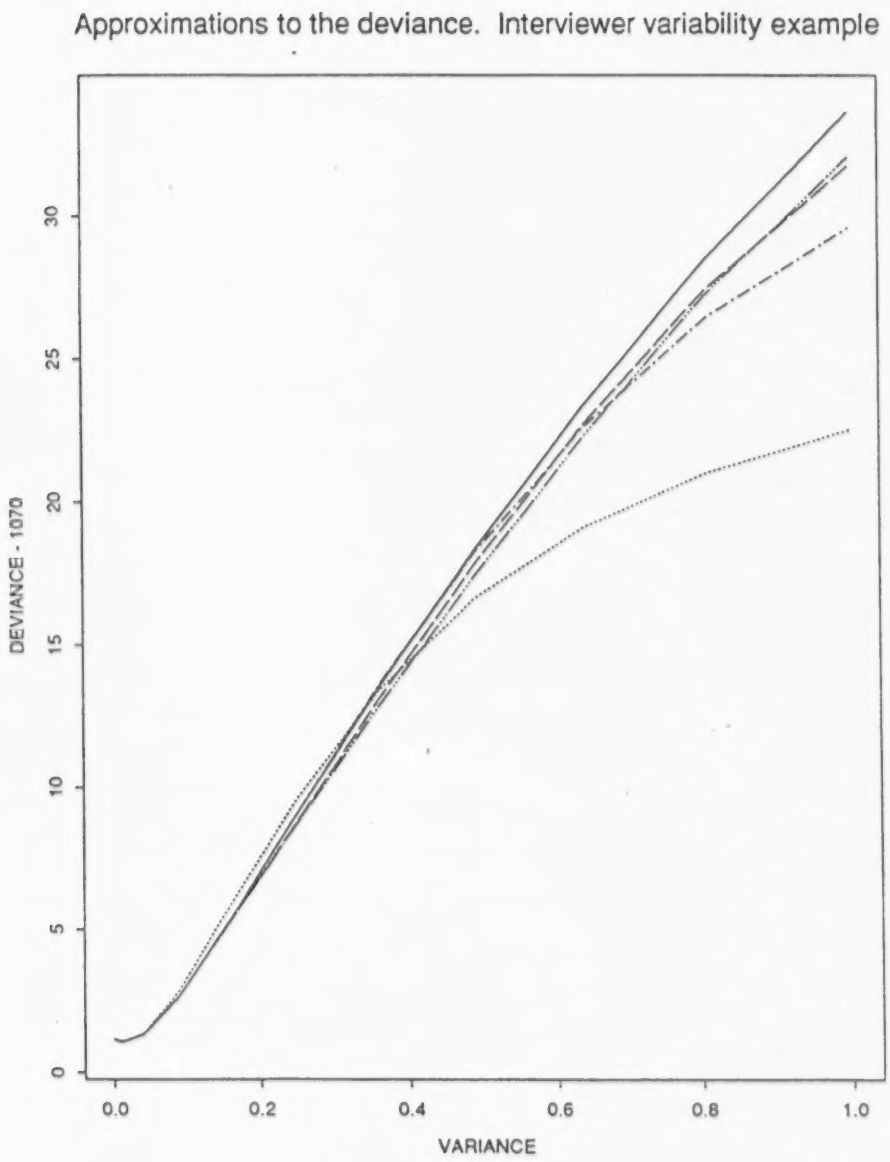

Figure 2: Profiles of the approximations to the –2 log-likelihood for the Interviewer variability data as a function of the variance σ^2.
The notation and methods of approximation are the same as in Figure 1.

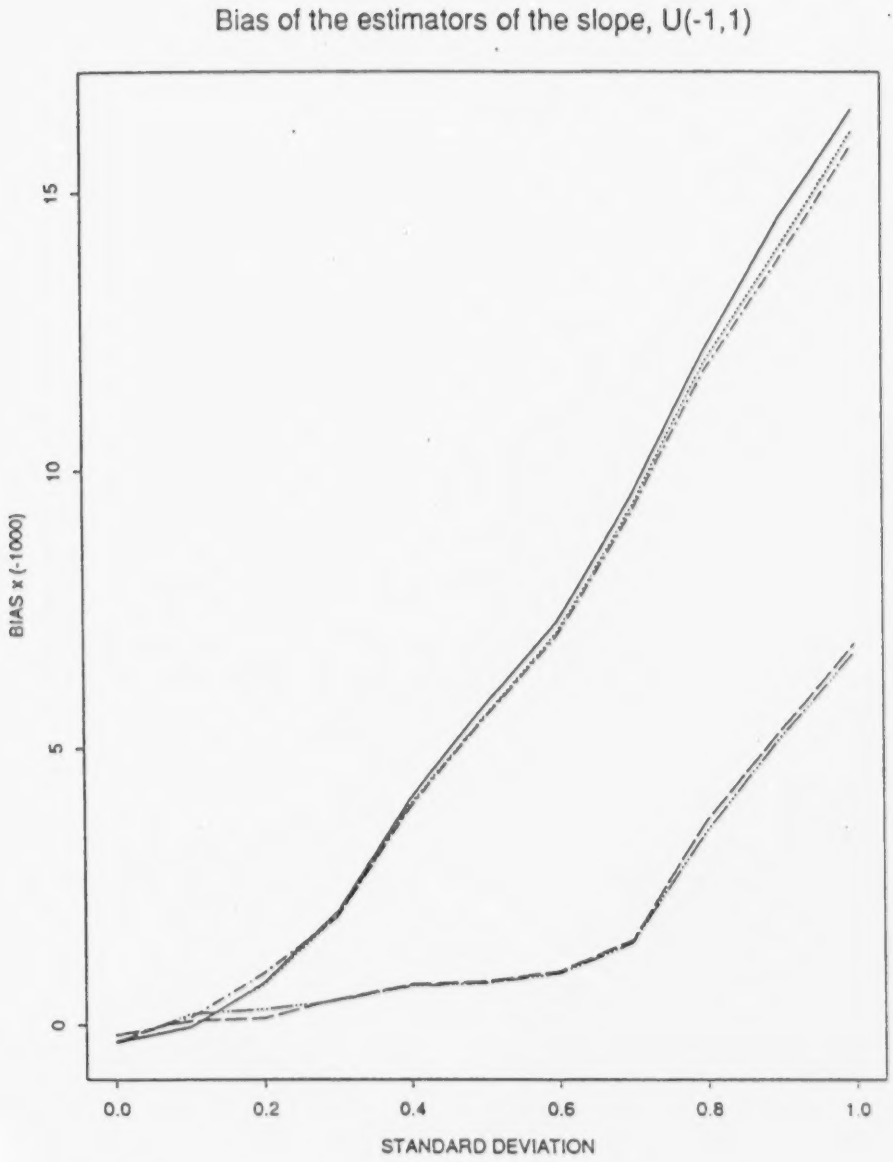

Figure 3: Bias of the estimators of the slope as a function of the standard deviation σ in the simulated data, U(-1,1) design.
The estimators are: —— GLS (generalized least squares); ..... AML (approximate maximum likelihood); —·—· RAML (approximate restricted maximum likelihood); —— ML9 (maximum likelihood using 9-point quadrature); —·—· REML9 (restricted maximum likelihood using 9-point quadrature. The scale of the vertical axis is in -.001 (most of the recorded biases are negative).

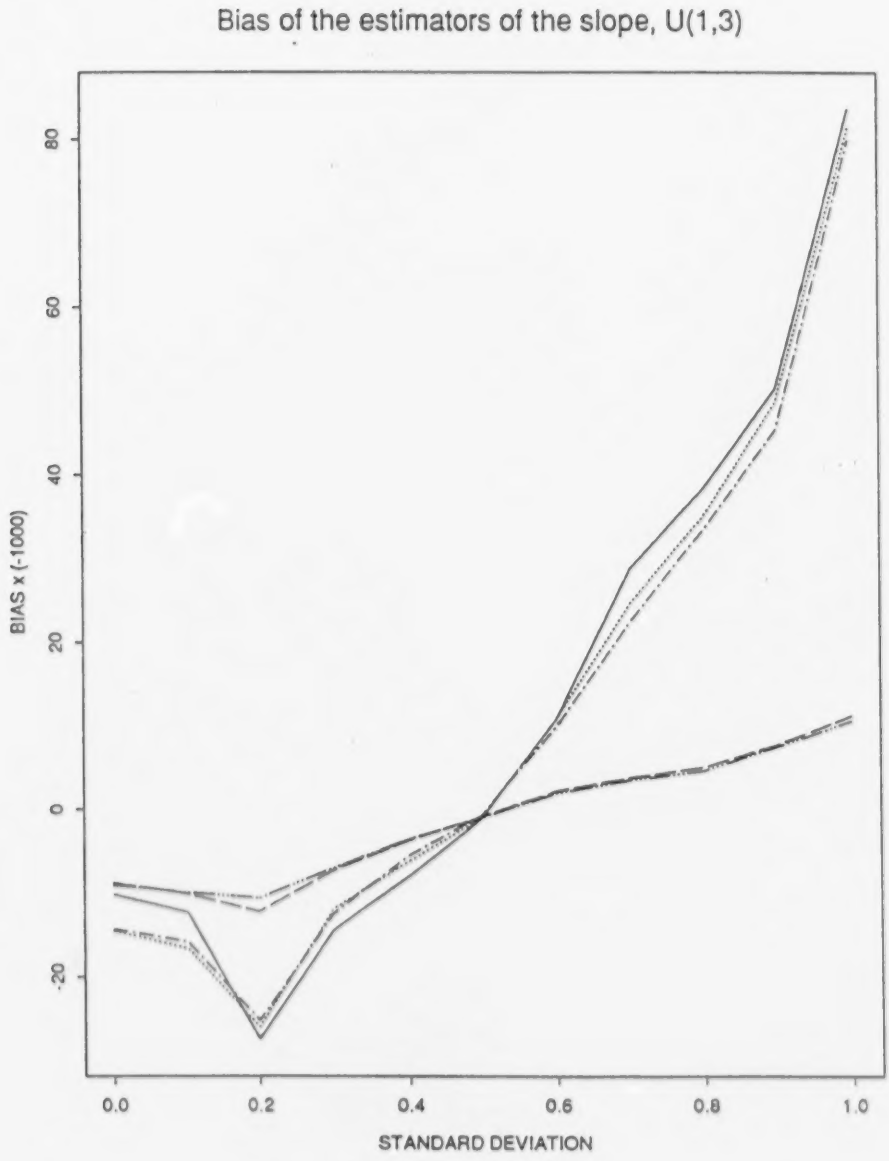

Figure 4: Bias of the estimators of the slope as a function of the standard deviation σ in the simulated data, U(1,3) design.
The estimators are GLS, AML, RAML, ML9 and REML9, as in Figure 3.

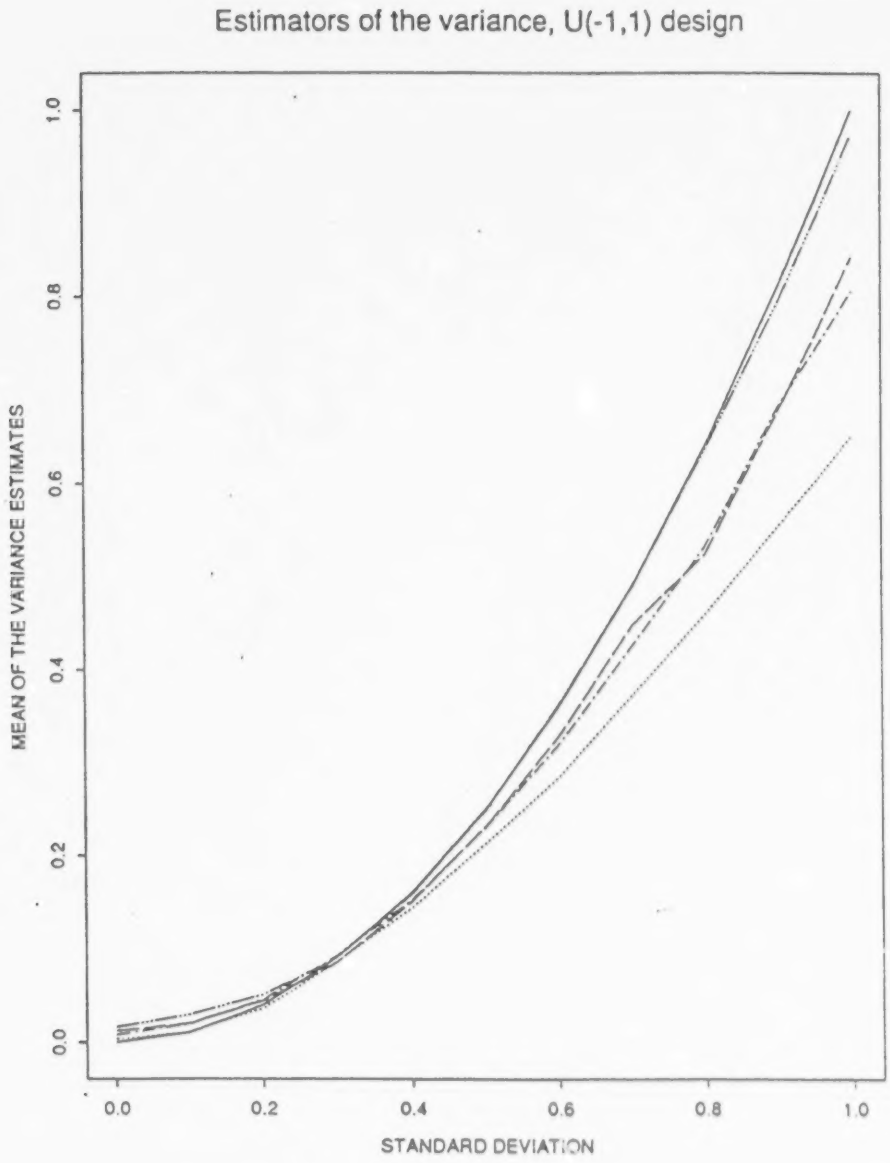

Figure 5: Comparison of the estimators of the variance σ^2; U(-1,1) design.
The estimators are: ..... AML (approximate maximum likelihood); —·—· RAML (approximate restricted maximum likelihood); —— ML9 (maximum likelihood using 9-point quadrature); ——·—— REML9 (restricted maximum likelihood using 9-point quadrature); ——— denotes the exact value of the variance.

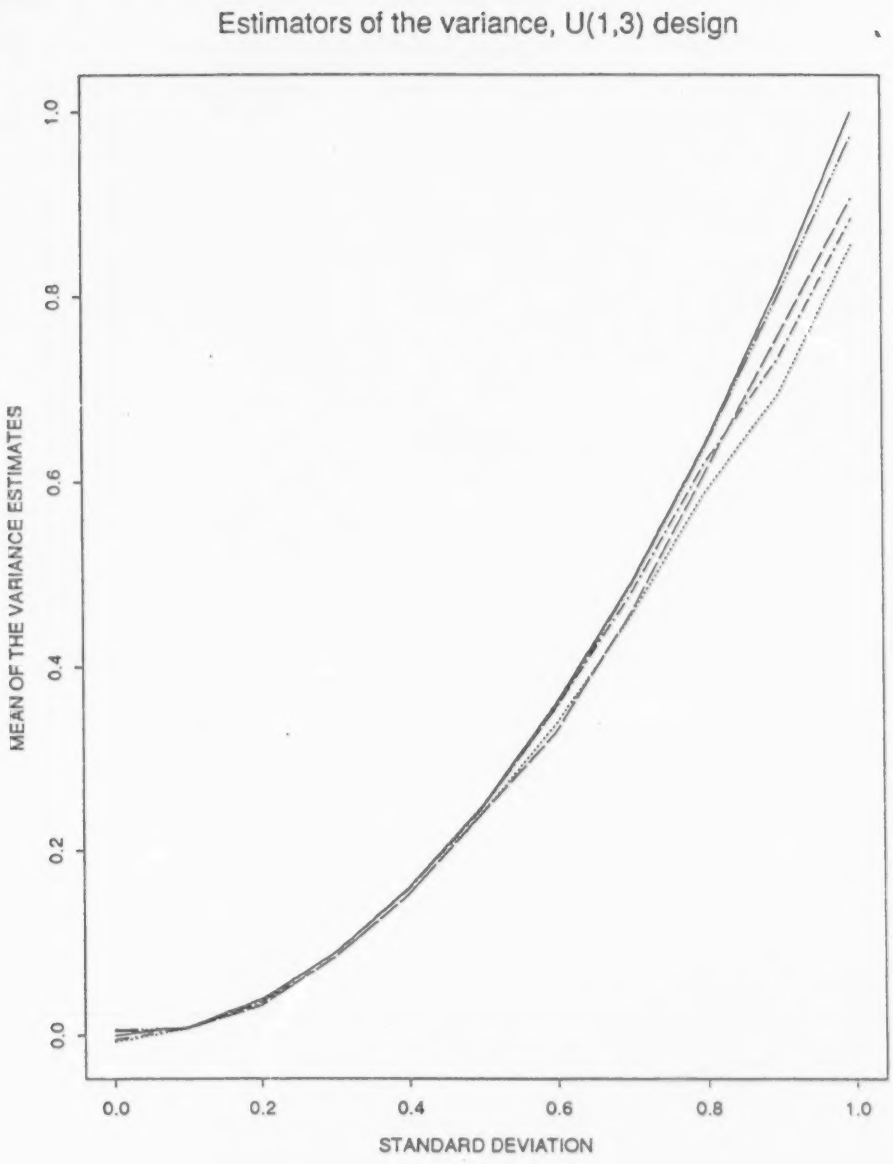

Figure 6: Comparison of the estimators of the variance σ^2; U(1,3) design. The notation is the same as in Figure 5.

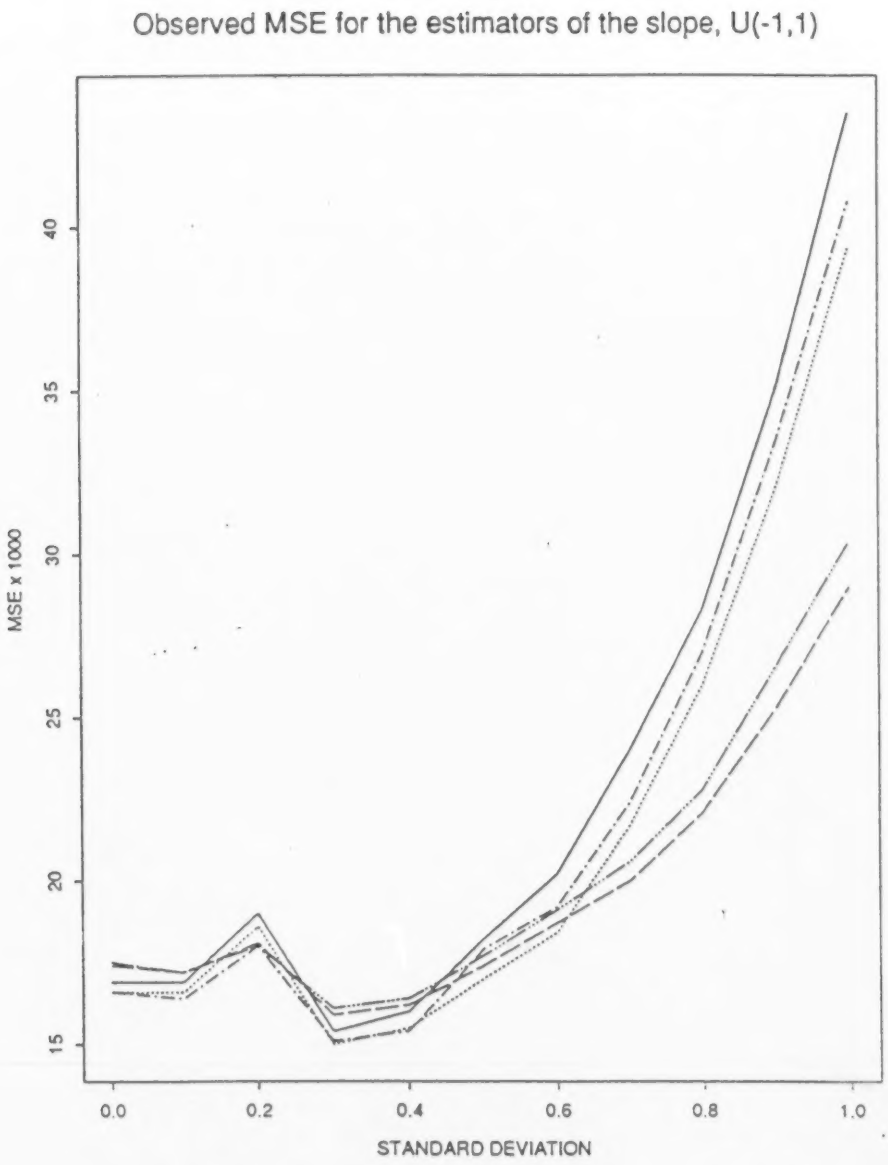

Figure 7: Mean squared error of the estimators of the slope as a function of the standard deviation σ in the simulated data, U(-1,1) design.
The estimators are: —— GLS; ..... AML; —·—· RAML; —— ML9; —···— REML9. The units on the vertical axis are 10^{-3}.

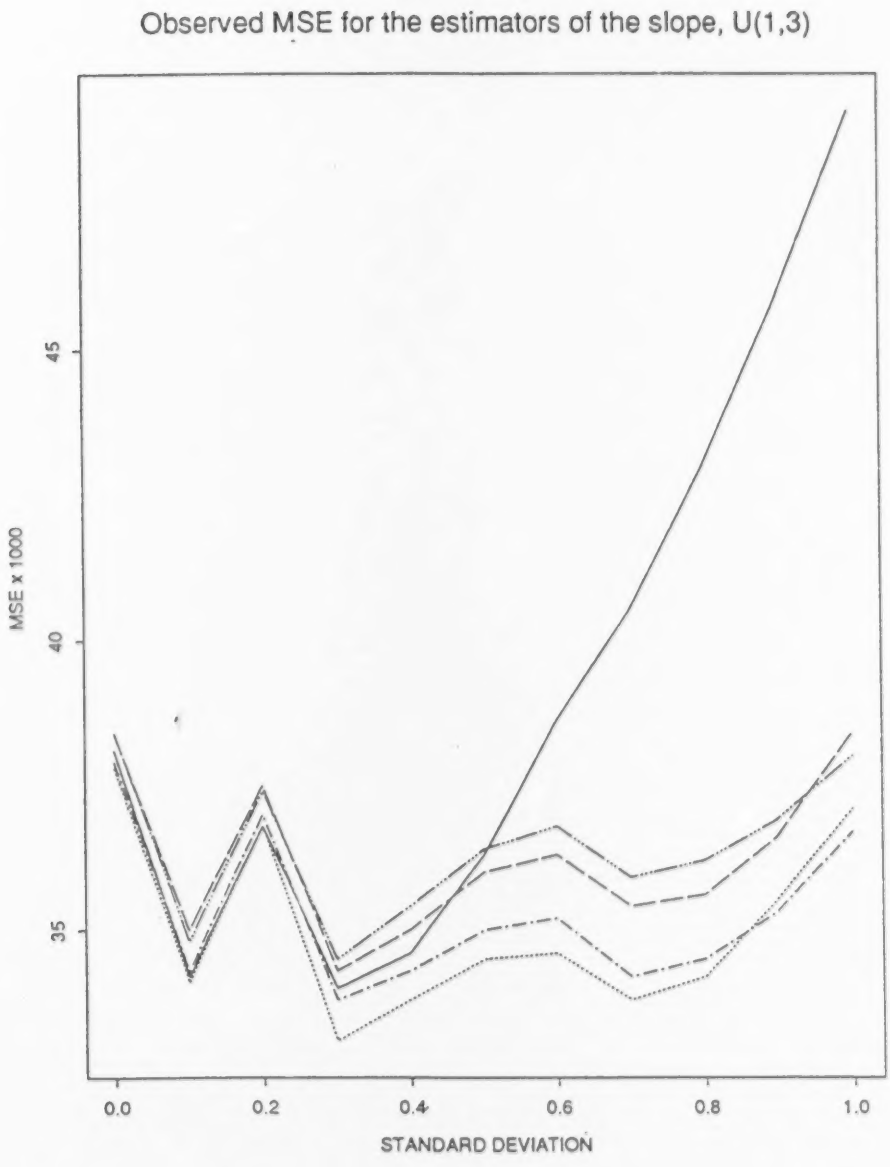

Figure 8: Mean squared error of the estimators of the slope as a function of the standard deviation σ in the simulated data, U(1,3) design.
The notation is the same as in Figure 7.

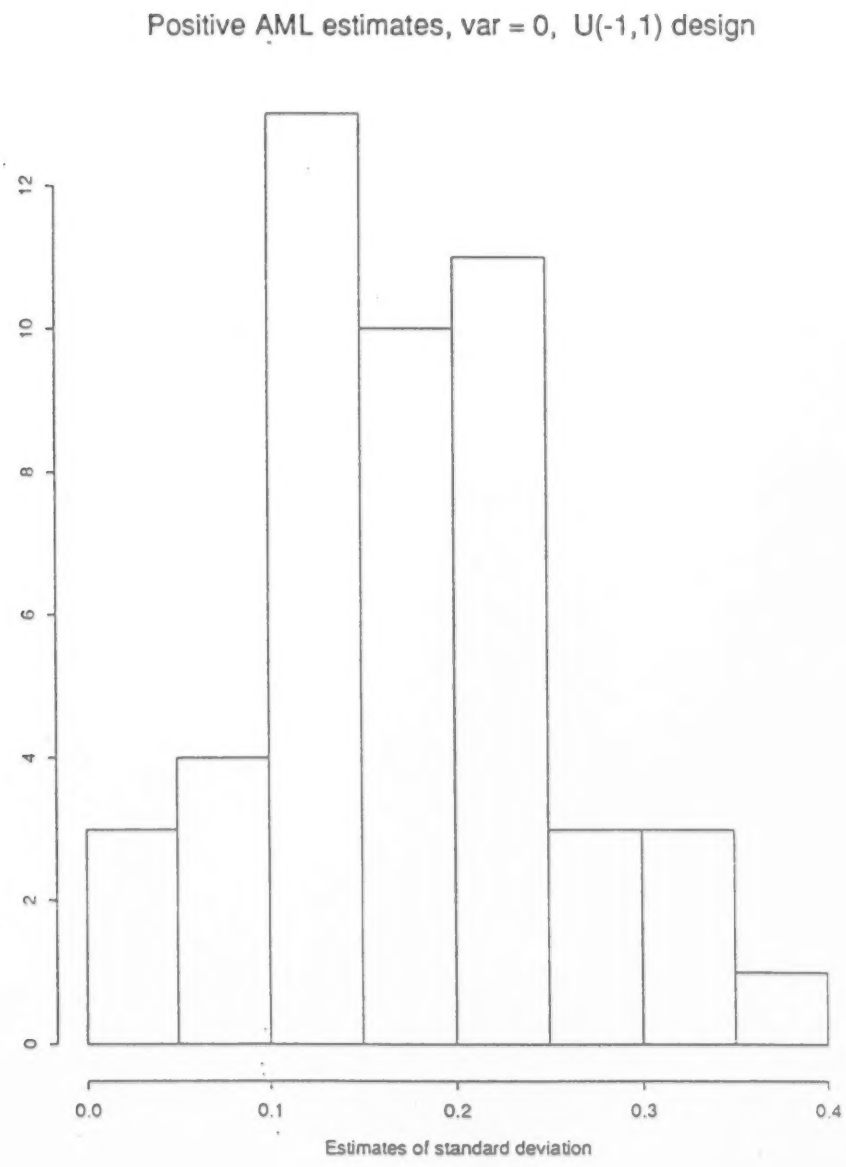

Figure 9: Distribution of the positive estimates of σ for the parameter value $\sigma = 0$ in simulated data, U(-1,1) design.
In the 100 simulations 48 positive estimates of the variance were obtained.

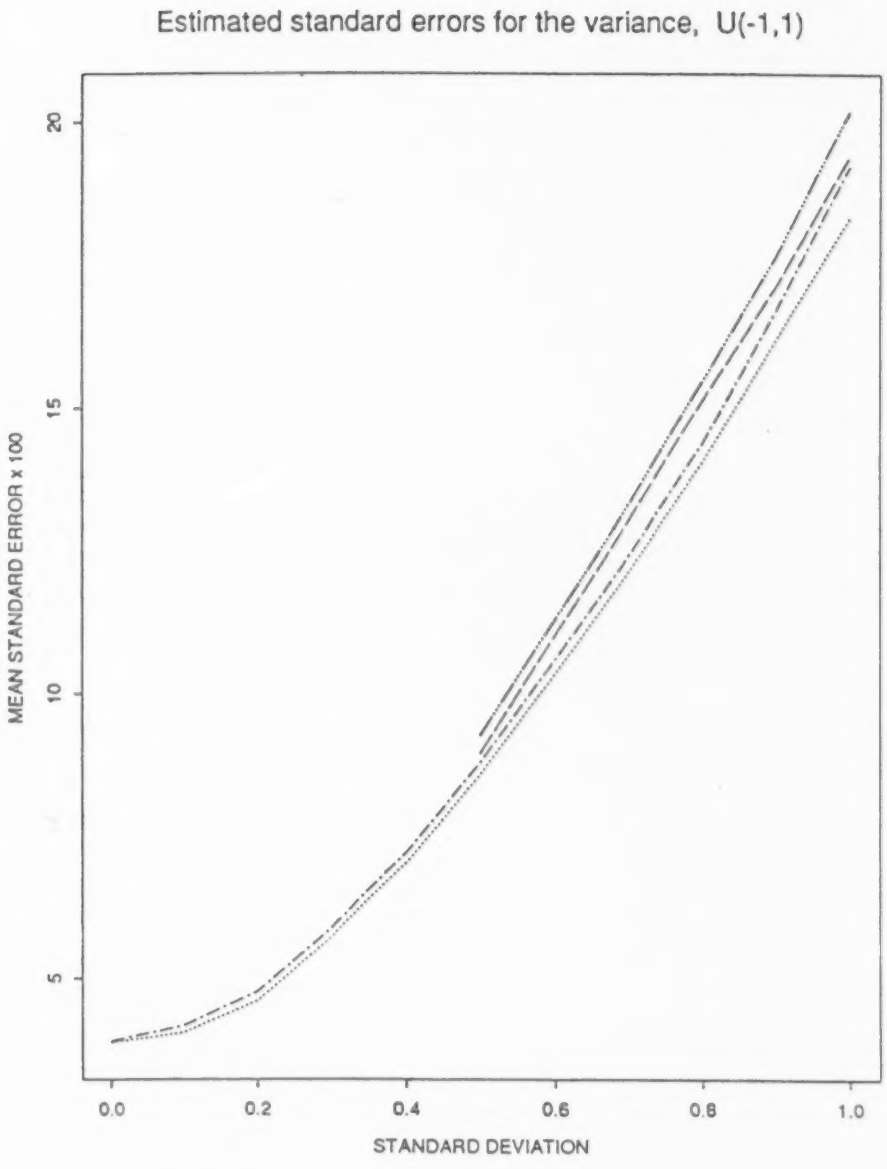

Figure 10: Estimated standard errors for the variance estimators for the simulated data, U(-1,1) design.
The methods of estimation are: ····· AML; —·— RAML; —— ML9; —···— REML9. The units on the vertical axis are 10^{-2}.

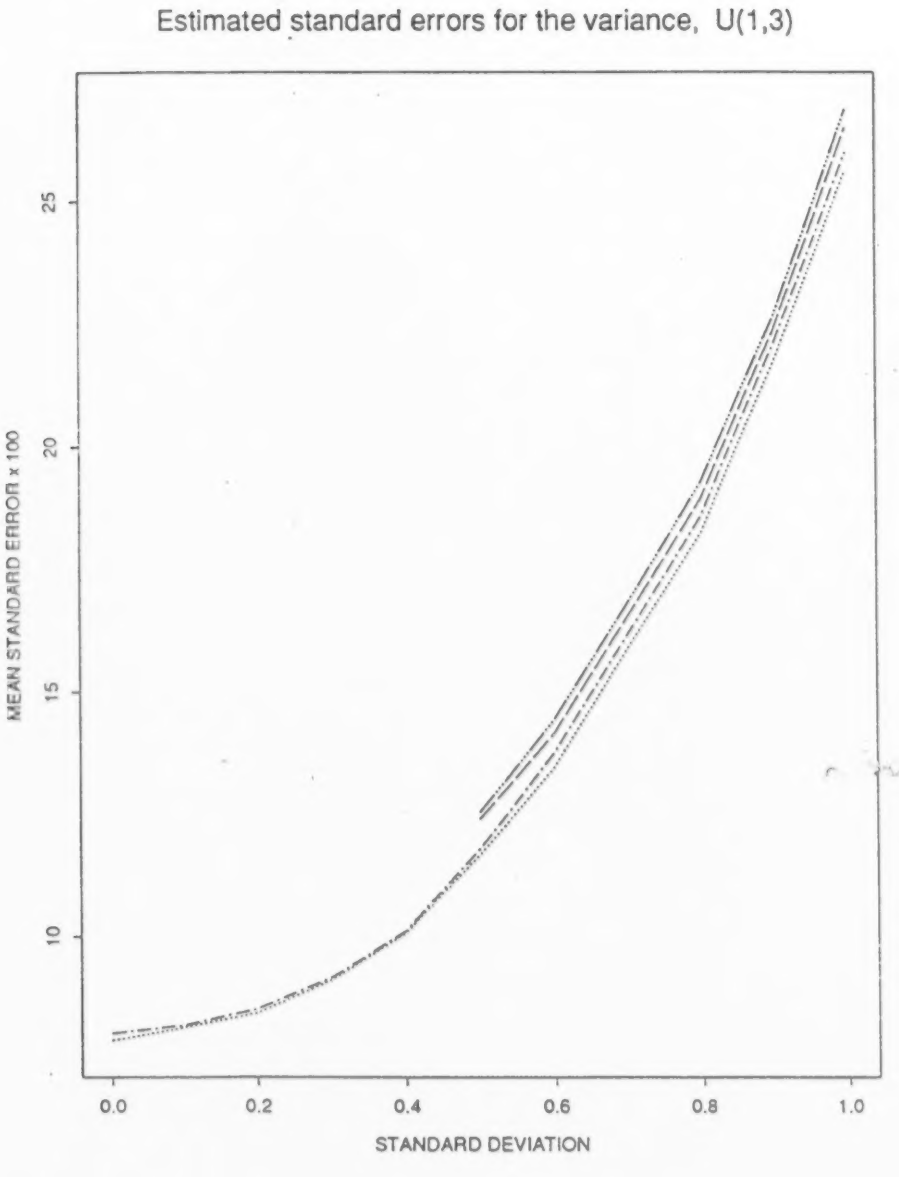

Figure 11: Estimated standard errors for the variance estimators for the simulated data, U(1,3) design.
The notation is the same as in Figure 10.

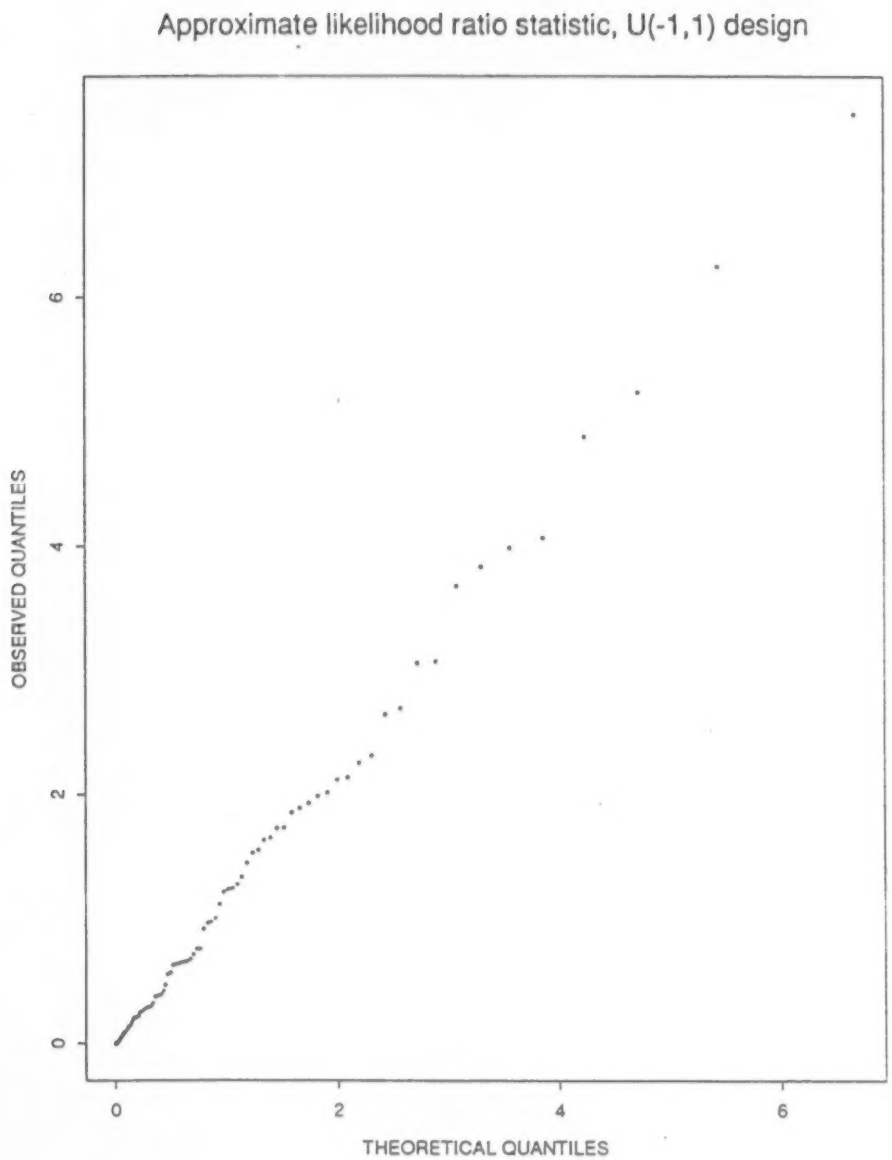

Figure 12: The qq-plot of the empirical null-distribution of the likelihood ratio statistic, against the χ^2_1 distribution, for the hypothesis of within-cluster independence ($\sigma^2 = 0$); the simulated data, U(-1,1) design.